# Prioritisation of Active Compounds that are Novel Scaffolds using a Data-Driven Approach


Hannah Jemi Patel

Linacre College
University of Oxford

A thesis submitted for the degree of
*Doctor of Philosophy*

Trinity Term 2018

To my sister Vicki,

As the most hard-working person I know it seems fitting to dedicate this thesis to you. I could never have imagined the amazing journey this DPhil would be and studying at Oxford has been without a doubt the most unbelievable experience. I hope that when you write your thesis in four years' time you feel the same way.

# Acknowledgments

First, I would like to say a massive thank you to all nine of my supervisors. Your support, guidance and continual effort have been invaluable, and I am so grateful. All of you have helped me to become a better scientist.

Second, I have to thank OPIG –the best research group in Oxford. Claire, thank you for your advice and reassurance and for showing me that being a scientist doesn't mean you can't make everything you do beautiful. Jin, thank you for sharing an excellent taste in music. Jaro, thank you so much for fixing my PyMOL... I've lost count of how many times. Eleanor, thank you for always listening to me rant about dieting, for always encouraging me to eat cake and for your infectious enthusiasm. Cristian, thank you for talking about sport constantly. Nick, thank you for your head-bobbing whilst your work - it was very entertaining and for helping me think of a name of my algorithm. Alex, thank you for letting me have half of your desk. Konrad, thank you for introducing me to python pickles – it was unbelievably helpful. Lyuba, thank you for being an inspirational scientist. Eoin, thank you for making me choke with laughter in group meetings. Susan, thank you for letting me complain about my problems and for having similar problems! Clare, thank you for your support, your constant willingness to have a chocomilk with me and for buying me that golden balloon– you have no idea how much that meant! Laura, I am so glad I met you, you will be a friend for life; thank you for always being there, for having a million breaks with me and for all the row chat.

To both of my school Chemistry teachers Dr. Bright and Mrs Macey. Thank you for encouraging me from a young age to pursue academia and study a PhD. Your enthusiasm and love of science and chemistry was infectious and I am so thankful to have been directed towards this.

Thank you to David Hagan who encouraged me to apply for Oxford, to apply to Linacre and for introducing me to the DTC. Without your knowledge and enthusiasm, I might never have picked this great college and this wonderful programme. I am so grateful.

A huge part of my life whilst completing this DPhil has been rowing. Linacre Boat Club, thank you for igniting my passion for rowing, and for the blades! Gemma, Gemma and Beth you have been the greatest friends and thank you for the many laughs shared, the support when we've had tough times and for the silverware we've won.

OUWLRC, you are the most wonderful club and thank you for making the two years I spent trialling so rewarding. I cannot wait to watch you win the Boat Race next year. To the 2017 Blue Boat, thank you for welcoming me so warmly when I came into the crew so late, for that BUCS final, and for the laughter. To the 2018 Blue Boat, thank you for giving me another chance to win, for the best row on the day and for being some of the most dedicated and impressive people I know. Laure, thank you for the many breakfasts at Gail's and for constantly encouraging me and rooting for me to make the Blue Boat. Maline, thank you for the great outings in the pair, for

always being there for me and for the many shared eyerolls. Andrew, thank you for telling me to go for it – I never would have thought being part of the Blue Boat was a possibility without you. Thank you to Jill for taking a chance on me in those early days and always being a great listener. Thank you to Clive for all your advice, and for shouting at me during that 2k. Thank you to Chris, for your guidance, for your coaching, for selecting me for two blue boat crews, and for making me seat race enough that I knew I earned it! Ellie, you are the most amazing person and I am so glad I met you. Thank you so much for being an amazing friend and making trialling as enjoyable as it could be. I won't make you uncomfortable by saying any more.

To Ben, thank you for putting up with me for the last four years. For cleaning the flat when I was tired, for listening to me moan when work wasn't going my way, for letting me cry when I thought I would never finish this thesis and for reminding me that it's okay to take time for yourself and relax.

Finally, to my family - the Patels. Where can I begin. You are all my inspiration. You all work harder than anyone else I have ever met, and every day I am inspired to work harder, to do better and to achieve because of all of you. Thank you for all your support and love, I love you all so much and I could not have done this without you.

# Abstract

A primary aim of drug discovery is to find novel molecules that are active against a target of therapeutic relevance. With high attrition rates and the increasing costs of each stage of drug development, there is an emphasis on making the right decisions in the early stages of drug discovery. Methods are needed that will aid selection of compounds as part of the hit-to-lead process that will avoid subjective decision-making and make use of the recent increase in experimental data available at the start of the drug discovery process.

I have developed the CRANkS algorithm to prioritise candidate compounds based on how novel the compounds are compared to known inhibitors of the target of interest. A prospective compound is compared to known binders of the target in terms of its predicted interactions with the protein, placement in the binding site, and chemical structure. Molecules are then scored based on the overlap of conformations with grids generated from protein-ligand X-ray crystal structures. The grids capture the spatial distribution of the chemistry of the ligands and protein-ligand interactions.

By using the calculated novelty scores, I hypothesised that the algorithm will prioritise both compounds that are active, but also compounds that are novel scaffolds. Scaffold-hopping is of particular importance in drug discovery for a better exploration of the activity space in order to find the most amenable lead compound. Using multi-objective optimisation, I found that the CRANkS scores could be used to select novel active scaffolds from datasets of active and inactive compounds.

Finally, I have investigated "active-guided" docking by combining the grids generated from active ligands, developed as part of CRANkS, with AutoDock – "AutoCRANkS". I found an improvement in discriminating between active and inactive compounds when using AutoCRANkS compared to AutoDock, indicating the promise of using protein-ligand structural data to guide docking.

# Declaration

I declare that no parts of this thesis or its research herein have been reproduced or accepted for another award or degree or diploma at any other university or learning institution. This thesis contains no other person's work except where stated in the text.

Hannah Jemi Patel
25th September 2018

# Contents

# List of Figures

## Chapter 1 Introduction



## Chapter 2 CRANkS Algorithm

## Chapter 3 CRANkS Algorithm II

## Chapter 4 Active-Guided Docking

## Chapter 5 Case Studies

# List of Tables

## Chapter 2 CRANkS Algorithm



## Chapter 3 CRANkS Algorithm II



## Chapter 4 Active-Guided Docking



## Chapter 5 Case Studies

# Chapter 1

# Introduction

## 1.1 Drug Discovery

A primary aim of drug discovery is to find novel compounds that are sufficiently active with a given therapeutic target (Medina-Franco *et al.*, 2014). Traditionally this follows a process that starts with preclinical research, beginning with target validation. A protein target is discovered that when inhibited or activated, causes a therapeutic effect on the disease or condition to be treated (Lipinski and Hopkins, 2004). Phenotypic screening is an alternative approach. This is followed by hit identification: compounds are found that have some activity with the target. These hits then need to be filtered and developed into lead compounds as part of hit-to-lead optimisation. This step forms the key focus of this thesis. It involves but is not limited to *de novo* molecular design: using computational methods to create libraries of molecules that are not already covered by existing patents. A subset of these molecules then need to be selected for synthesis in terms of predicted potency or other physicochemical properties, and these properties are then optimised as part of lead optimisation. The lead compounds can then be taken forward into clinical trials (Lipinski and Hopkins, 2004). An overview of the drug discovery pipeline is shown in Figure 1.1.


Figure 1.1 The process of drug discovery: libraries of molecules are screened against a target to find hits. The hits are modified, and some are developed into lead compounds. The leads are then optimised before clinical trials. Hit-to-lead and lead optimisation are dominated by subjective decision making are outlined in purple. As the process proceeds the money and time required for each step increases.

As the process of drug discovery proceeds, not only does the money and time invested increase but the cost and time required for each step also increases. The total cost for the development of a new drug can range from $161 million to $1.8 billion (Morgan *et al.,* 2012), and Phase III clinical trials can make up 90% of the total costs of research and development of a drug (Roy, 2012). With the growing cost of research and development and the higher attrition rates of drug candidates at the final stages, there is an increasing emphasis on ensuring the best possible candidate or candidates are taken forward from preclinical research to clinical trials.

The increasing cost of the discovery of a drug can be attributed to three fundamental reasons. The first is increasing complexity due to polypharmacology (Reddy and Zhang, 2013; Hopkins, 2008), whereby drugs regulate the activity of several targets. Secondly, drugs are increasingly failing at the clinical stages, the most expensive of the stages of the process, due to lack of activity of the drug against the target or due to toxicity (Cook *et al.,* 2014). This is indicative of taking forward an inappropriate lead compound, and of problems in the hit-to-lead and lead optimisation process (Hann, 2011). Thirdly, it has become increasingly difficult for drugs to be approved, and

although the process of having a drug approved by the Food and Drug Administration (FDA) has been reduced in time and cost, the process is still more expensive and difficult than in previous decades (Ciociola *et al.,* 2014).

The failure of drugs at the clinical stages has increased the emphasis on ensuring the best possible candidate is taken forward from preclinical research. Subjective-decision making has dominated hit-to-lead and lead optimisation. Medicinal chemists select compounds for follow-up based on synthetic intuition, existing biases or experiences. Studies have shown that different medicinal chemists use the same characteristics to select compounds, but different weighting and preferences lead to a lack of agreement between medicinal chemists (Kutchukian *et al.,* 2012). Additionally, there is little consensus about which properties of molecules are undesirable. This selection can have a huge effect on the success of the drug discovery process. Therefore, methods that reduce or avoid this subjective decision-making could significantly improve the process.

### *1.1.1 Hit-to-lead and Lead Optimisation*

Marketed drugs have been identified to be highly similar to the lead compounds from which they have been optimised (Proudfoot, 2002). It is therefore crucial for the most amenable compounds to drug development be taken forward at the hit-to-lead and lead optimisation stages, and for a number of potential leads to be available, to maximise the chance of finding best-in-class medicines. The lead optimisation process involves optimising a number of properties required for an active compound to become a drug.

The “drug-likeness” of the compound refers to the similarity of characteristics or properties of the compound to other successful drugs, and it is assumed that molecules with sufficient drug-likeness are more likely to have the required attributes to become an approved drug. These properties include activity, the molecule needs to sufficiently modulate the target, but also selectivity – the activity of the molecule against other biomolecules or proteins that are not targets. Off-target activity can often produce unwanted side-effects (Smyth and Collins, 2009). Toxicity is another property requiring optimisation: pre-clinical and clinical toxicity are the cause of approximately one third of attrition rates (Kola and Landis, 2004). Pharmacokinetic and pharmacodynamic properties affect drug administration and dosage and are additional properties that cannot be neglected (Meibohm and Derendorf, 1997). Figure 1.2 shows a representation of this multi-factor problem and the trajectory of a typical lead optimisation program. The optimisation of one property can lead to the detriment of another property.

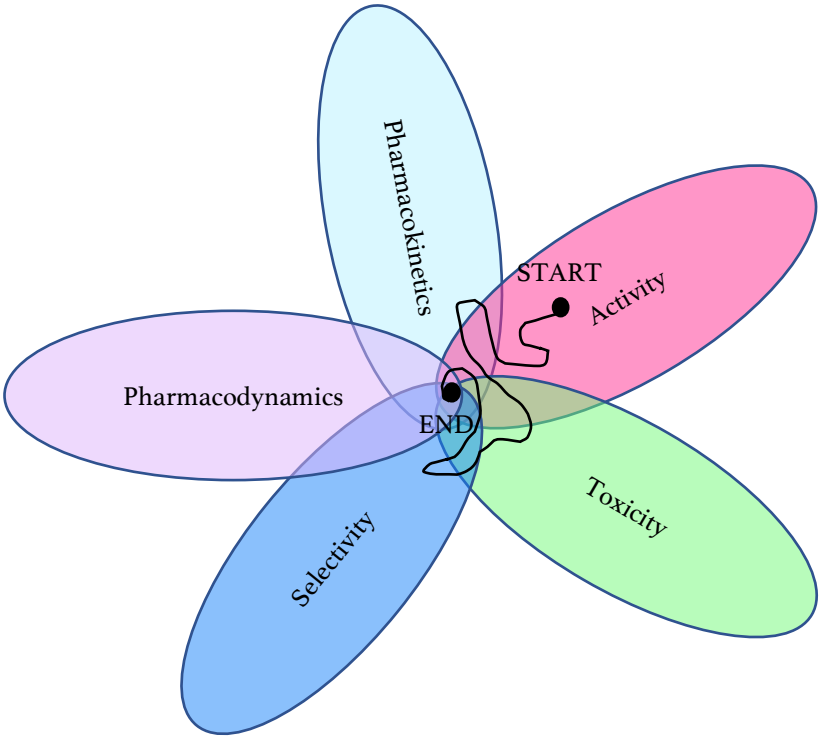



Figure 1.2 Representation of the multi-factor aspect of the lead optimisation process. All five of these properties must be optimised for a successful lead-like compound and typically as one is optimised there is a detriment to at least one of the other properties.

Traditionally, medicinal chemistry has been used for the lead optimisation process following an iteration of the cycle shown in Figure 1.3. First compounds are tested to determine which properties require improvement. The results are evaluated and used to decide upon a series of follow-up molecules that could improve the properties in question. This evaluation is often based on synthetic intuition and subjective-decision making. Although less subjective methods are available, such as the early Topliss method (Topliss, 1977) that uses a decision tree to generate aryl substituents for a series of analogues in question, there is a clear need for workflows that optimise not only potency, but the other properties required for a drug-like molecule.

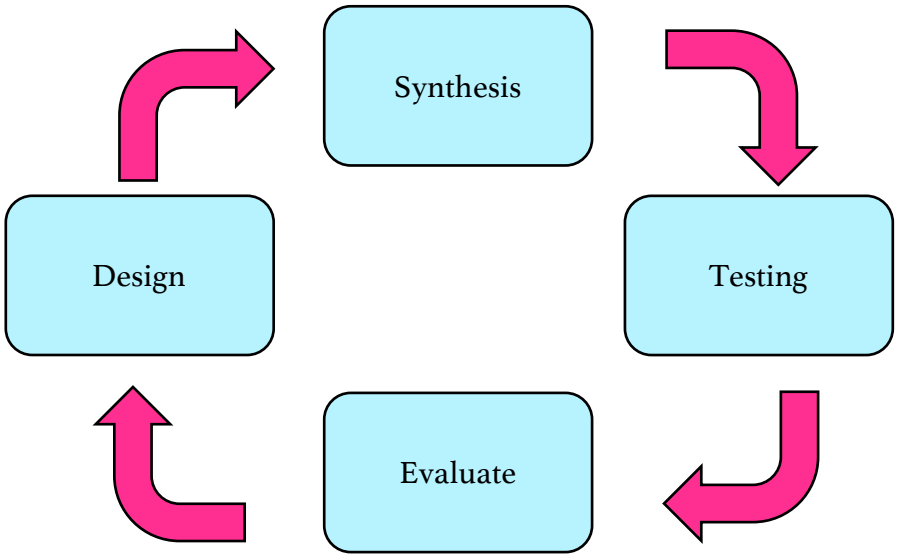



Figure 1.3 The traditional medicinal chemistry methodology adopted as part of hit-to-lead and lead optimisation. Testing of compounds identifies which of the properties need to be improvement by evaluation of the testing data. This influences the design of new compounds which should yield improvement of these properties. These new compounds are synthesised. The process repeats until compounds with desired properties are obtained.

## 1.2 *De novo* Molecular Design

In some ways drug discovery can be described as a sampling question (Schneider and Schneider, 2016). The process involves selecting a set of compounds with the potential to be drug-like from an almost infinitely large set of possible compounds. The number of possible unique drug-like molecules has been estimated to be $10^{60}$-$10^{100}$ (Dobson, 2004; Lipinski and Hopkins, 2004). Computational work offers the potential to search through a larger amount of this space, more quickly, than traditional laboratory work. However, despite the amount of computational power now available, it is still unfeasible to search this space exhaustively.

Figure 1.4 shows an illustration of chemical space adapted from "Development II" by Escher (Dobson, 2004). Within chemical space there are several activity islands representing desirable compounds with the required properties to be a successful drug candidate. Each of the activity islands contains structurally different sets of molecules i.e. different scaffolds but can be considered to interact with the target in the same way. To move from one island to another – to "scaffold-hop" the chemist must navigate through areas containing compounds with less amenable properties or pharmacological profiles (Schneider and Fechner, 2005).

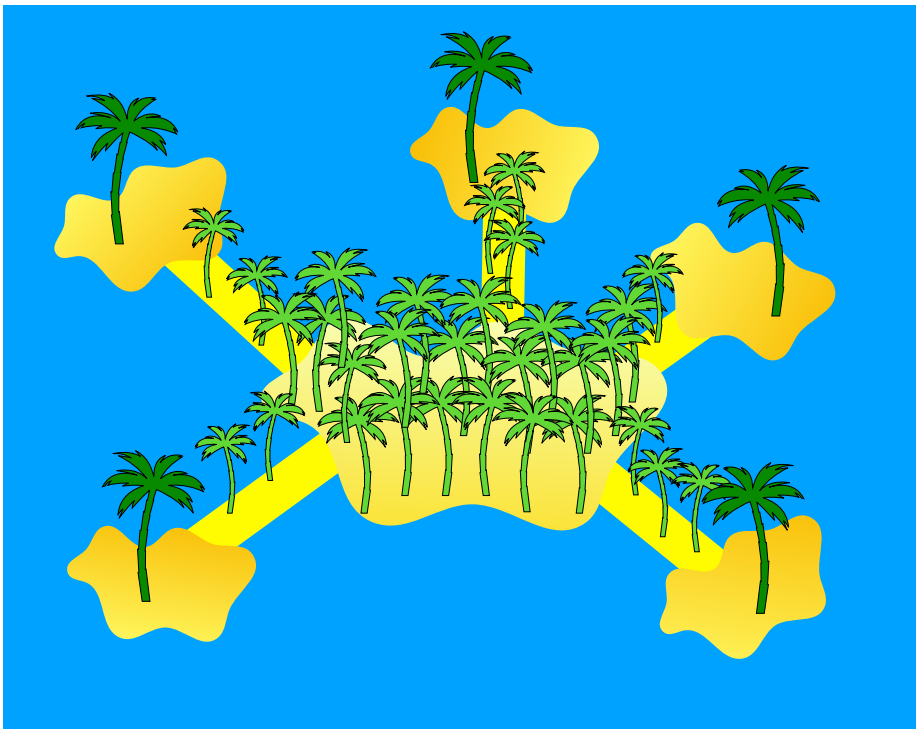


Figure 1.4. An illustration of drug-like chemical space adapted from "Development II" by Escher (Schneider and Fechner, 2005). Each of the islands contains a drug-like compound with the desired properties to become a successful drug candidate, illustrated by a dark green palm tree. It is impossible to move directly between islands as the islands are separated by sea. To move between islands i.e. to scaffold-hop, areas containing less desirable compounds illustrated by light green palm trees must be traversed.

*De novo* molecular design looks to design novel compounds from scratch that are active against the required target with the desired drug-like properties (Schneider, 2012). This requires navigation of this chemical space. *De novo* design consists of three parts each with separate issues to overcome: compound generation, scoring of the compounds and optimisation of those compounds. This effectively covers the majority of *in silico* medicinal chemistry. Compound generation can yield libraries of millions of potential compounds that will then be sub-selected to libraries of hundreds to thousands. However, even this can be too many compounds to synthesise. Compounds can then be further scored and analysed to select candidate compounds for synthesis and testing by biophysical assays. This scoring and selection step is the main focus of this thesis.

Ideally *de novo* molecular design methods look to facilitate scaffold-hopping. Scaffold-hopping refers to searching for compounds that have similar potency but a different chemical structure (Schneider *et al.,* 1999). This is central to modern medicinal chemistry for a number of reasons. Firstly, scaffold-hopping can improve the properties of a lead compound to make it a more amenable drug candidate, for example by changing to a more soluble scaffold, or improving the pharmacokinetic properties. Secondly, smaller changes to the structure can lead to a novel structure that is then patentable: so-called patent-busting (Bohm *et al.*, 2004). Computational algorithms are useful for systematic searching of different scaffolds for scaffold-hopping and multiple methods have been developed for this, frequently using pharmacophore models (Hu *et al.,* 2016).

### *1.2.1 Compound Generation*

*De novo* molecular design generally begins with the generation of novel compounds. These are generated by the grouping of building blocks. The building blocks are typically atoms or fragments (Schneider and Fechner, 2005). Initial structure generation tools were originally based around fragments and the use of the target structure (Schneider and Fechner, 2005). An initial fragment would be positioned within the binding site to maximise interactions with the target. The fragment would then be grown or linked with further fragments from a library, such as RECAP (Lewell *et al*., 1998), to optimise for further interactions in the binding site. These methods have seen a resurgence due to the popularity and success of fragment-based drug discovery (Section 1.6). However, this only allows for a local search and could not be used to navigate a larger chemical space to facilitate scaffold-hopping.

The generation of structures using computational methods is generally considered to be adequate (Schneider and Schneider, 2016). Both ligand and receptor-based approaches have been developed that allow the generation of drug-like novel compounds, and the available computing power means that many compound structures can be quickly calculated (Schneider, 2013; Korb *et al.,* 2014; Reymond, 2015). Reaction-based molecular design yields synthetically tractable molecules and the DOGS (Design Of Genuine Structures Software) molecule assembly tool has been shown to lead to the discovery of novel chemical compounds i.e. scaffold-hop, by using 25000 commercially available fragments recombined for rapid exploration of chemical space and selecting compounds based on pharmacophore similarity

(Hartenfeller *et al.,* 2012). The tricky and somewhat unsolved part of the process is selecting the most amenable candidates from the library of enumerated compounds (Schneider and Schneider, 2016). The scoring and selection of compounds is discussed in the following sections.

## 1.3 Ligand-Based Scoring

Scoring compounds can be split into two categories: ligand-based and structure-based. Structure-based scoring uses a structure of the target and is discussed in Section 1.4. Ligand-based scoring is typically used when no target structure is available but a compound that binds to the target – a hit – is known (Dean *et al.*, 2004). There are many ligand-based scoring methods that have been used with success in drug discovery. Here I will discuss two of the most commonly used methods – molecular similarity and quantitative-structure activity relationships (QSAR).

### *1.3.1 Molecular Similarity*

Much of ligand-based scoring can be related to the similarity property principle. This states that similar chemical compounds are more likely to have similar properties, in this case the property being the binding affinity of the compound (Johnson and Maggiora, 1990). An example of when this principle holds true is shown in Figure 1.5. These compounds are morphine and codeine, each of which have a similar structure and have similar properties – both drugs are opiates that can be used for pain-relief

and are highly addictive. Of course, there are numerous cases when this heuristic fails. Two compounds that are similar but have different binding affinity are known as activity cliffs (Maggiora, 2006; Stumpfe *et al.*, 2012; Stumpfe *et al.*, 2014). In some cases, simply the addition of a methyl group can have a vast effect on the potency of a compound and this is known as the magic methyl effect.

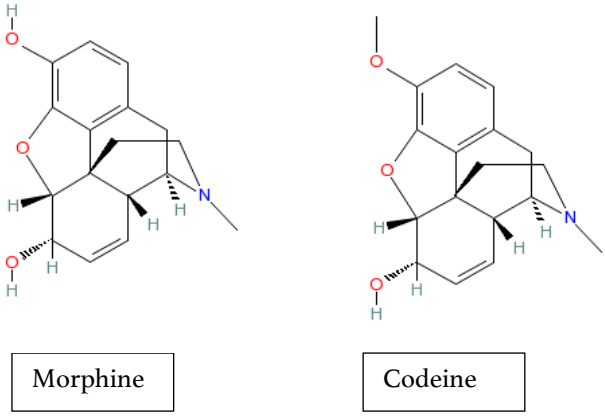



Figure 1.5. Morphine and codeine are classic examples of the similarity property principle. The chemical structure of the drugs is very similar, and this yields similar therapeutic properties. Both are opiates and can be used for pain-relief and are also highly addictive.

The concept of molecular similarity is subjective and there are many different ways to measure the similarity between two compounds. Computational calculations of similarity can be used to apply objectivity to molecular similarity which will allow results to be both unbiased and invariant over time (Willett *et al.,* 1998). To calculate molecular similarity between two compounds there are two fundamental parts: a representation of the compound that details the relevant features (molecular characterisation) for similarity and a function to calculate similarity (a similarity coefficient) (Maggiora *et al.*, 2014).

There are also multiple types of similarity each of which have different uses, and this is summarised in Figure 1.6, with the typical descriptors used for each. The terms

"chemical" similarity and "molecular" similarity are often used interchangeably but in truth are slightly different concepts. Chemical similarity refers to the similarity of molecules in terms of molecular properties and physicochemical characteristics. Chemical similarity makes use of 1-dimensional descriptors, such as logP or molecular weight, to describe the molecules. These do not effectively capture any structural features or chemical connectivity within the molecule. It can also use synthetic information or functional groups (Maggiora *et al.,* 2014).

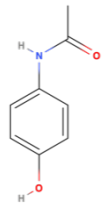
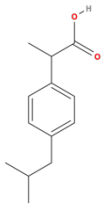
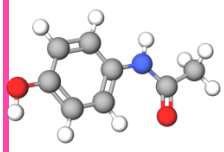
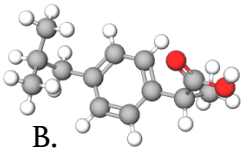
| Similarity | Descriptor Examples | Dimensionality | Illustration | |
|---|---|---|---|---|
| Chemical Similarity | Molecular Weight, LogP, Number of Rotatable Bonds, Number of Aromatic Rings | 1D | A. 1 aromatic ring, MW = 151 g/mol | B. 1 aromatic ring, MW = 206 g/mol |
| Molecular Similarity | 2D Fingerprints | 2D | A.  | B.  |
| Molecular Similarity | 3D Fingerprints, ROCS descriptors | 3D |  A. |  B. |

Figure 1.6. Illustrations of chemical similarity and molecular similarity and descriptors for each.

Molecular similarity, on the other hand, refers to the similarity of the structural components of compounds and how atoms are linked together. This can be in 2-

dimensions or 3-dimensions, depending on the molecular descriptor used. 3-dimensional descriptors can be built over an ensemble of generated ligand conformations as the active conformation is usually unknown. This is rational as the shape of the ligand has a significant impact on binding. However, 3D descriptors can suffer due to the large amount of information encoded leading to a significant amount of noise, or due to inaccurate conformer generation. An example is the shape- based ROCS descriptor, which uses atom-centred Gaussian functions to describe molecules (Grant *et al.*, 1996). 3-dimensional descriptors can also be highly dependent on how the molecules are superimposed, unless non-superpositional descriptors such as Ultrafast Shape Recognition (USR) or ElectroShape are used (Ballester and Richards, 2007; Armstrong *et al.,* 2010).

The most commonly used molecular descriptors are 2D methods which capture the local chemical connectivity of the molecule. Although this does not capture the 3D shape, the amount of information is reduced. This causes a reduction of noise and fast generation of descriptors. 2D descriptors are most commonly converted into fingerprints where molecules are described by bit strings (Duan *et al.*, 2010). Each bit represents the presence or absence of a chemical feature or substructure generated from an atom connectivity table. The chemical features or substructures encoded are defined by the fingerprinting method. For example, pharmacophore-based fingerprints would encode the presence of pharmacophoric features using a bit (McGregor *et al.,* 1999). A pharmacophore can be defined as "an ensemble of steric and electronic features that is necessary to ensure the optimal supramolecular interactions with a specific biological target and to trigger (or block) its biological response" (Wermuth *et al.*, 1998).

The most popular molecular similarity coefficient is the Tanimoto similarity (Bajusz *et al.,* 2015; Tanimoto, 1957), also known as the Jaccard index, *JI*:

$$JI = \frac{c}{a+b-c} \tag{1.1}$$

where *a* and *b* are the counts of the non-zero bits of the fingerprints of molecules A and B, and *c* is the count of the non-zero bits that the molecules share in common. Other similarity metrics have also been developed and there is contention on which is the most suitable. However, the Tanimoto coefficient remains the most popular and has proven to produce similar results to other more recently developed coefficients (Bajusz *et al.,* 2015).

Molecular similarity coefficients are dependent on the fingerprint used to characterise the molecule (Bender, 2010; Duan *et al*., 2010). Additionally, the coefficient does not give any indication of which parts of the molecule are similar or why the molecules are similar, making it difficult for the user to interpret the score or make informed decisions when selecting molecules for follow- up. There is also no information of the implications of the differences or similarities between the molecules or the ability of the molecule to bind to the target.

### *1.3.2 Quantitative-Structure Activity Relationships*

Quantitative-structure activity relationships (QSAR) are based on the principle that the physiological effect of a compound is related to its chemical structure (Brown and Fraser, 1868; Brown 2016). A mathematical model is generated that uses independent variables consisting of descriptors of compounds and usually outputs the activity of a compound against a given target. QSAR can also be used in 3D by making use of 3-dimensional descriptors of the compound. The typical process for the use of QSAR is shown in Figure 1.7.

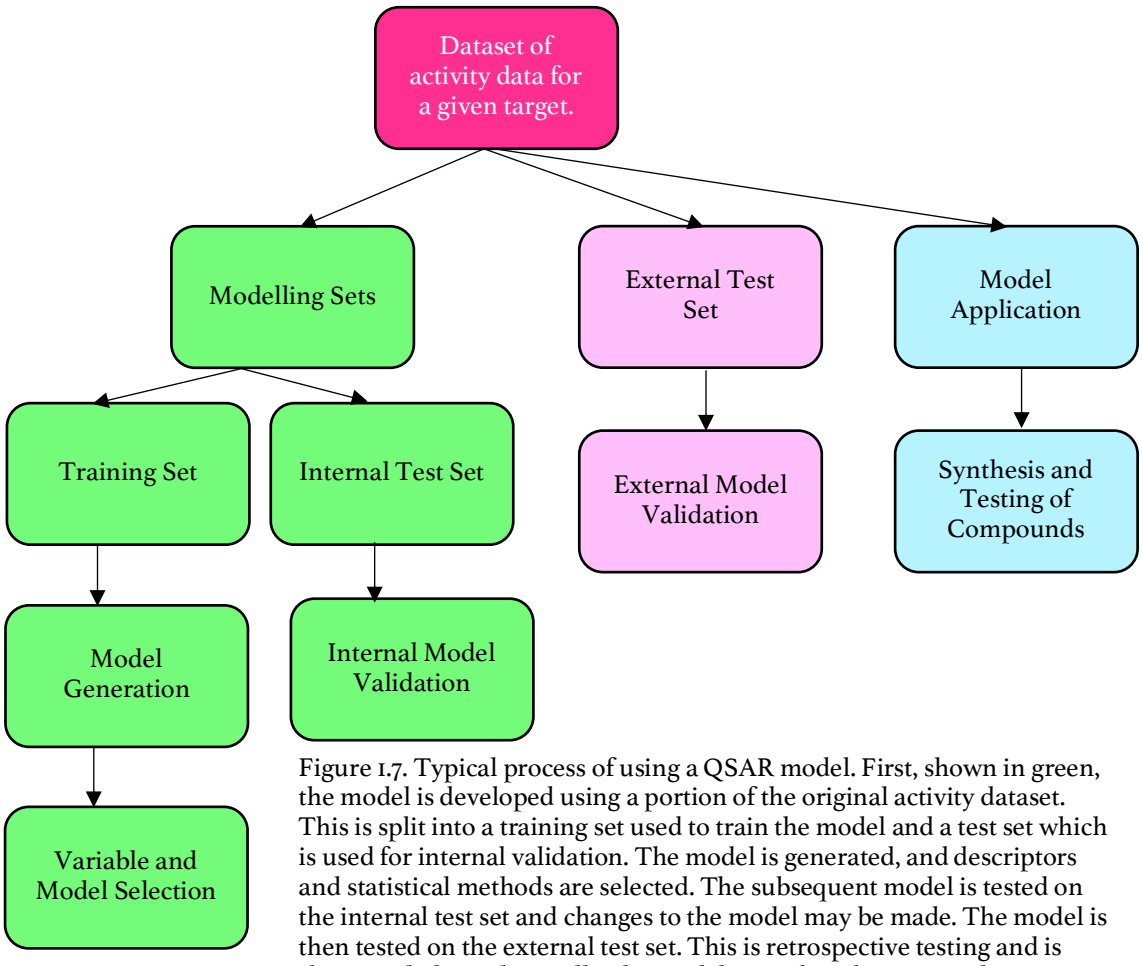



Figure 1.7. Typical process of using a QSAR model. First, shown in green, the model is developed using a portion of the original activity dataset. This is split into a training set used to train the model and a test set which is used for internal validation. The model is generated, and descriptors and statistical methods are selected. The subsequent model is tested on the internal test set and changes to the model may be made. The model is then tested on the external test set. This is retrospective testing and is shown in light pink. Finally, the model is used, and compounds are synthesised and tested based on the predictions. This is prospective testing and shown in light blue. QSAR figure 1 is adapted from Brown, 2016.

One of the most important steps in building a QSAR model is the selection of molecular descriptors that will be used to generate the model (Gedeck *et al.,* 2006), for the success of the model but also the interpretability. Algorithms such as the

"Greedy Forward Selection" method selects descriptors by iteratively adding descriptors and determining whether the quality of the model is improved or not (Brown, 2016). However, this is a long routine and success was found by Paul Labute by hand selecting descriptors that can be used in different scenarios from a set that are easily computed (Labute, 2000). Genetic algorithms are the most popular way of selecting descriptors (Brown, 2016).

There have been numerous successful cases of using this method but on the other hand there are also well-published issues (Cherkasov *et al.,* 2014). Simple modelling strategies, such as linear regression, are unable to capture the nuances or complex aspects of the data. Conversely, complex modelling strategies that use many descriptors or use more sophisticated statistical practices often require a large amount of training data, are liable to overfitting (Hawkins *et al.,* 2004), or can be difficult to understand (Manchester and Czermiński, 2008). A key problem with the use of QSAR models are the lack of measures of the reliability of the model, and prospective testing of a QSAR model is now necessary for a journal publication (Tropsha *et al.,* 2003; Brown 2016).

## 1.4 Structure-Based Drug Design

Structure-based drug design has the potential to be used to find the most amenable lead compounds for pharmaceutical targets (Bleicher *et al.,* 2003). The process makes use of the structure of the protein target to guide hit-to-lead and lead optimisation. Protein structures can be determined by electron microscopy, NMR spectroscopy, or most commonly X-ray crystallography. There have been multiple instances when using this approach has led to an accelerated discovery of a drug. One example is vemurafenib which gained approval from the FDA a mere six years after the project began (Erlanson *et al.,* 2016).

The process for structure-based drug design is illustrated in Figure 1.8 adapted from a review by Anderson (Anderson, 2003). First a target structure is required. This can be an experimentally-determined or an active compound can be docked into a protein structure to provide a protein-ligand complex. Either a medicinal chemist or a computational algorithm can be used to propose modifications to the compound to generate lead compounds. Next, the lead is evaluated either by scoring or visualisation of the interactions formed. A subset of the lead compounds can then be selected for synthesis. The limitations of the synthetic tractability, reactivity or insolubility of *de novo* compounds generated and selected by computational methods means that this is often performed by a medicinal chemist and can be subjective and influenced by the chemist's personal bias and experience of synthetic chemistry (Hartenfeller, 2012). The recent developments in X-ray crystallography discussed below have caused a resurgence in the use of structure-based drug design in drug discovery.

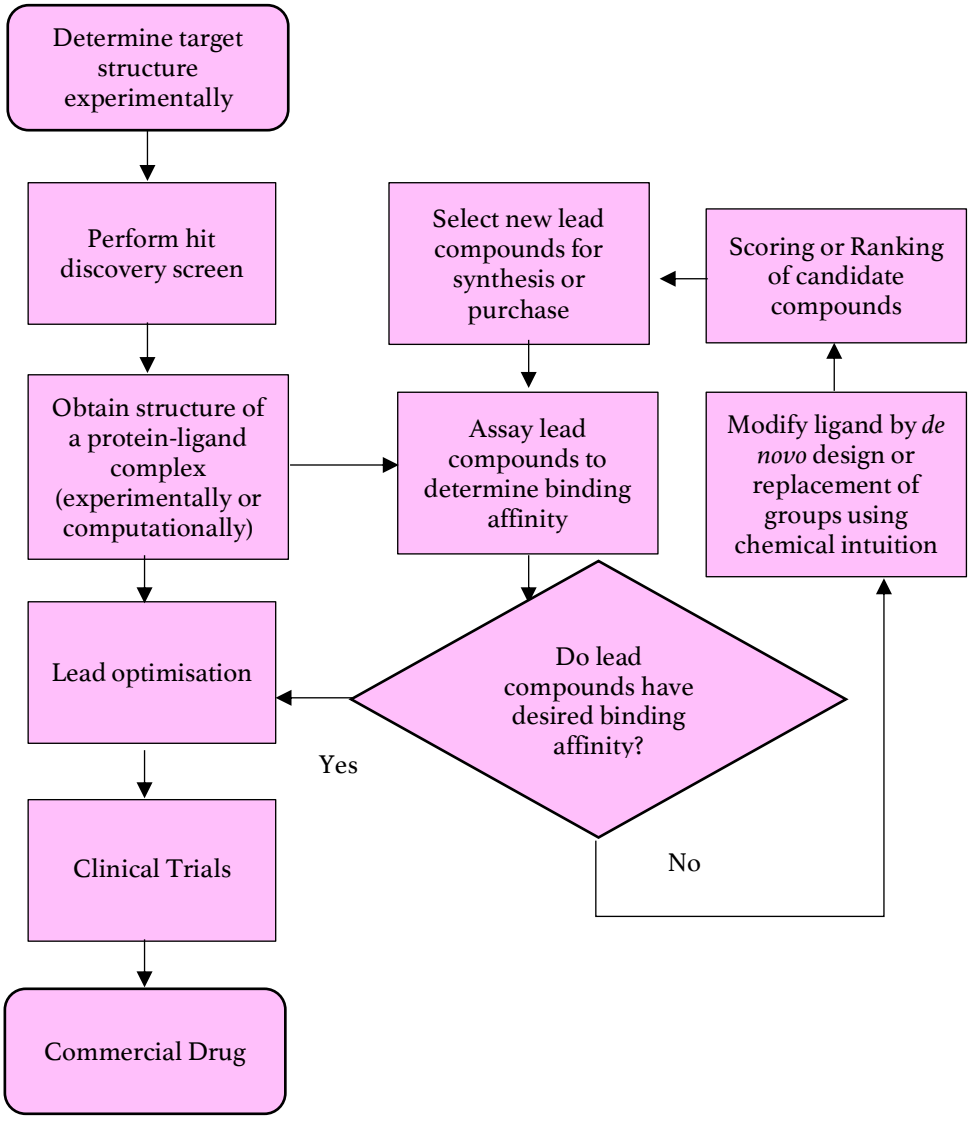



Figure 1.8. Schematic of the process of structure-based drug discovery adapted from Anderson, 2003.

## *1.4.1 X-ray Crystallography*

Initially X-ray crystallography could not be used as a primary screening method due to the time-consuming and low-throughput nature of the process. However recent advances in developing high-throughput workflows for crystallography (Winter and McAuley, 2011; Badger, 2012) and co-crystallisation of compounds with the protein (Patel *et al.,* 2014) mean that X-ray crystallography can now be used as a primary screening tool yielding structural information about active candidate compounds against a target.

X-ray crystallography benefits from low false positive rates as well as high sensitivity as compounds are able to be soaked at a much higher concentration than traditional solution-based assays. However, limitations of the method include a sizeable false negative rate – the binding of candidate compounds could disrupt the crystal lattice by inducing a change in protein structure.

X-ray crystallography works by directing a beam of X-ray radiation at the crystal structure of a molecule, or molecules, to obtain a diffraction pattern. This pattern is the Fourier transform of the electron density of the molecule in question, and so can be used to determine the 3D structure of the molecules. Only the amplitude of the diffracted waves can be detected. The phase cannot be collected and must be estimated. This modelling of the positions is usually achieved by molecular replacement which iteratively estimates the coordinates of atoms by calculating the agreement of the derived model with the data until a threshold is met (Abergel, 2013). A crystal, a repeating lattice of the molecule to be tested, is required for a strong enough diffraction to be detected.

There are a number of descriptors that are used to describe the quality of both the data collected and the subsequent crystal model. The resolution is the smallest separation of atom planes that can be discerned and is usually reported in Ångstroms. For protein-ligand interactions and the conformation of a bound ligand to be detected clearly and without ambiguity, a resolution of better than 2.5 Å is typically required. The R-factor describes the quality of the subsequent modelled structure and measures the agreement of the model with the X-ray diffraction data (IUCR, 2015). An R-factor lower than or approximately 10 times the resolution can be considered sufficient. The R-factor considers atoms used in the construction of the model in its calculation and so is liable to overfitting. Consequently, the free R-factor was developed that only uses 5% of the model in its calculations that was not used in the iterations of improving the agreement of the model with the data (Brünger, 1997).

Two more descriptors are important for structure-based drug design and the interpretation of a ligand bound to a protein. The first is the B-factor which is calculated per atom and measures the square of the deviation of the atom from the mean coordinates of the atom. This can indicate the strength of binding of regions of a ligand, as regions that are not bound tightly will have more flexibility and consequently higher B-factors. The second is occupancy. The occupancy of an atom refers to the proportion of molecules within the crystal found with that particular atom. Low occupancy atoms (less than 0.8) will have a weaker diffraction pattern and consequently modelling the structure of these atoms can be complicated and should be treated with caution.

### *1.4.2 Molecular Interactions*

The binding of a drug to a target relies upon the favourable interactions between them. The essence of structure-based drug design is that, by making use of the structure of a protein target, the interactions formed by a candidate compound can be optimised (Bissantz *et al.,* 2010). The first *de novo* molecular design strategies used the derivation of interaction sites to influence the design of compounds (Schneider and Fechner, 2005). HSITE calculated the hydrogen bond acceptors and donors available in the receptor yielding a map of areas that small-molecule atoms or donors can occupy to form potential hydrogen bonds (Danziger and Dean., 1989). Later methods expanded the interaction types available to first include hydrophobics before expanding to covalent bonding and ionic interactions with metals. However, these methods did not allow for sufficient exploration of the chemical space.

Calculation and subsequent visualisation of the interactions formed between candidate compounds or ligands is used frequently in hit-to-lead and lead optimisation. The computer can highlight interactions to the medicinal chemist that perhaps would otherwise be overlooked, allowing for a more unbiased approach to the interpretation of the interactions formed by a candidate compound. Although well-known interactions such as hydrogen-bonding and π-stacking can be recognised by eye, other weak interactions could be overlooked such as halogen bonds. Tools such as the Protein-Ligand Interaction Profiler, PLIP (Salentin *et al.,* 2015), and Arpeggio (Jubb *et al.,* 2017) calculate and allow the visualisation of protein-ligand interactions for drug discovery. Arpeggio was recently used to aid the development of a 0.6 micromolar inhibitor of a target enzyme for tuberculosis (Trapero *et al.,* 2018). However, interpretation and visualisation still require a user to

select compounds and does not exclude bias or intuition from this cherry-picking of candidate compounds.

### *1.4.3 Docking*

Docking is a method that attempts to accurately predict the binding affinity and binding pose of a candidate compound with a target protein structure. Candidate compounds are "docked" – conformers are generated for the compound and translated and re-orientated to fit within the target binding site and these conformations are scored to give an approximate binding energy for the compound (Kitchen *et al.,* 2004). Docking has frequently been used for hit identification as part of virtual screening but is also utilized during lead optimisation (Kitchen *et al.,* 2004). Variations of the structure of the lead can be quickly tested and the binding pose evaluated. Docking has also been used to predict the metabolism of drugs (Weaver and Gleeson, 2008; Chadwick and Segall, 2010).

Although docking has been used successfully to discover and develop numerous drugs over the last 30 years there are still a number of challenges faced by the method that prevent the method from fulfilling its full potential to aid drug discovery. (Wang and Zu, 2016). These can be split into three main factors: the inaccuracy of scoring functions, the disregard of water molecules in the binding pocket, and the inconsistency of accuracy of the protein-ligand structures generated.

Many scoring functions have been developed to attempt to efficiently and accurately calculate the binding affinity between a candidate compound and a corresponding

protein target (Jain, 2006; Huang *et al.,* 2010). However, the correlation achieved by these functions is approximately 0.7 for the most accurate functions available (Huang *et al.,* 2010), and despite the calibration of scoring functions with experimental data, or knowledge-based functions that are derived from available data, the generated scores are not the binding free energy. There are a number of alternative approaches to docking available that can accurately calculate the binding affinity of a ligand such as free-energy perturbation and thermodynamic integration (Reddy *et al.,* 2014). However, the computational demands of these methods prevent their use in docking as a high-throughput calculation of vast amounts of conformations of compounds are required. The main causes of the inaccuracy of docking scoring functions are detailed below.

Interactions known to make contributions to binding are frequently excluded from scoring functions such as halogen bonding (Ren *et al.,* 2014). Guanidine-arginine interactions are not included in any functions (Yang *et al.,* 2015). Classical scoring functions consider interactions to be additive to the overall affinity, however this is not always the case. For example, the strength of halogen bonding can be influenced by a number of interactions including water bridges (Ren *et al.,* 2014). Additionally, there may be interactions involved in the protein-ligand complex that are currently unknown, and this will also contribute to the inaccuracy of the scoring function (Wang and Zu, 2016).

Despite calibration on experimental data there is inherent inaccuracy with docking scoring functions as "one size does not fit all" (Ross *et al*., 2013). Using statistical learning theory and information theory, Ross and co-workers found that target-

specific models are likely to be more accurate than functions that aim to be applicable to any target.

Another contribution to the inaccuracy of docking is the treatment of water (Wang and Zu, 2016). Routinely many docking protocols begin by deleting all water molecules from the binding site. This is clearly a very different system to the actual environment of protein-ligand binding. Water molecules are removed from the docking process because the algorithms do not allow for the displacement of the solvent by the ligand. Although a number of methods have been developed to attempt to calculate the contributions of the water molecules to the energetics of the binding process this problem still has yet to be solved (Bodnarchuk, 2016).

Finally, the treatment, or lack thereof, of protein flexibility adds to the inaccuracy of scoring functions but also adds error to the conformer generation step of the docking. In most cases the flexibility of the target is completely ignored, and few docking program can consistently handle the perturbation of the binding pocket or changes to the structure caused by ligand binding (Lexa and Carlson, 2012; Wang and Zu 2016). Although it is accepted that in general conformer generation is adequately accurate there are numerous cases when the docked pose is very different from the structure experimentally determined, usually due to ligand flexibility or flexibility of the protein (Erickson *et al.,* 2004). Indeed, even if the correct pose is generated the inaccuracy of scoring functions means that the pose could be removed and not proposed as one of the final docked conformations.

## 1.5 The Potency Trap

Hit-to-lead and lead optimisation has increasingly shifted away from prioritising compounds based on predicting potency. This is in part due to the low accuracy of binding affinity prediction currently achieved by scoring functions: correlation with binding affinity is approximately 0.7 for the best scoring functions (Hang *et al.,* 2010). However, it is also due to the "potency trap", whereby molecules are selected with high potency that is achieved by increased lipophilicity (Hann, 2011). This makes it difficult to optimise the other characteristics that make a good drug candidate, especially oral bioavailability. This tendency is referred to as "molecular obesity" (Hann, 2011). Therefore, alternative methods have been developed to select compounds without solely using the predicted potency of the compound. Some examples of these are discussed below.

### *1.5.1 Molecular Diversity*

Selecting molecules based on molecular diversity is one method that has been developed to avoid prioritising potency too early on in the drug discovery process. Chemical space can be regarded as infinitely big in terms of potential molecules and estimates for the number of possible active small molecules range from $10^{60}$ to $10^{100}$ (Dobson, 2004; Lipinski and Hopkins 2004). However, for a compound to be able to be used as a drug it must have certain properties associated with absorption, toxicity and stability in the body, as well as potency and specificity (Lipinski and Hopkins, 2004). Thus, the space available when searching for therapeutic agents is significantly smaller than the chemical space of all organic molecules. This makes

exploration of this space important for finding the most suitable drug for a target. At the hit-to-lead stage it is therefore considered to be important to have as high a diversity as possible to explore chemical space as broadly as possible. This should maximise the chance of finding a suitable lead (Galloway *et al.*, 2010).

Although there is contention about how molecular diversity is defined (Roth, 2005), a number of computational methods have been developed to pick a diverse subset of molecules from a database (Willet, 1999). The subset should contain compounds with sufficient structural variation to represent the molecular diversity of the whole dataset. The diverse subset should also be small enough to be synthesised within the required time. A description of molecular diversity can quantify the choice of these compounds.

Molecular diversity of a dataset of compounds can be calculated in a three-step process (Willet, 2014). Firstly, the molecules must be characterised in terms of their chemical and structural properties. Secondly the similarity between molecules must be calculated by comparing these characteristics. Finally, the diversity of the subset can be summarised by calculating a diversity index from the molecular similarities.

Molecular characterisation has been found to significantly affect the results for molecular diversity calculations (Bender, 2010; Duan *et al.*, 2010). It is hard to measure directly how accurate molecular descriptors are when used to calculate molecular similarity. This is due to a difference in the definition of molecular similarity: a variation with which molecules are interpreted by medicinal chemists and a single chemist when shown the same molecules repeatedly has been found

(Lajiness *et al.*, 2004). Bioactivity coverage was recently used as an external measure for molecular descriptors (Koutsoukas *et al.*, 2014). This found that 2D fingerprinting methods show high coverage of bioactivity space and recommended the use of a combination of 2D fingerprints for selecting a diverse subset of molecules from a database (Koutsoukas *et al.*, 2014).

Many computational algorithms have been developed to select the most diverse subset of compounds from a large database (Martin, 2001). These either first identify similar groups of molecules and use this to form dissimilar subsets, or directly identify a diverse subset using dissimilarity-based algorithms (Gillet, 2011).

Cluster-based and cell-based subset selection methods fall into the first category. The cluster-based method groups molecules based on molecular similarity and selects either a single molecule or a number of molecules from each cluster (Varin *et al.*, 2009). If the clusters are dissimilar enough this should produce a diverse subset of molecules. Cell-based methods plot molecules on a low-dimensional grid defined by chemical properties (Lewis *et al.,* 1997). The molecules fall within different cells depending on the molecules' combination of chemical properties. A diverse subset is then selected by picking molecules from different cells.

Dissimilarity-based algorithms pick an initial molecule and build up the subset by subsequently picking either the molecule that is most dissimilar or from a selection of molecules that are more dissimilar than a defined cut-off (Snarey *et al.*, 1998). These algorithms use diversity indices to quantify the similarity between molecules as discussed in Section 1.3.1.

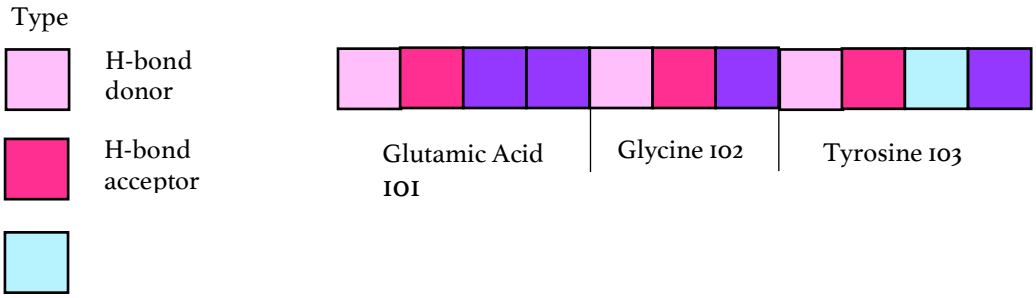
These methods do not usually account for synthetic feasibility. A medicinal chemist will select molecules from this subset to synthesise based on their intuition on what is tractable synthetically, and the resources available. However, this unaided selection could pick the most similar molecules from the subset which could compromise the diversity of the subset relative to the original dataset, introducing significant bias.

More recently the focus has shifted away from molecular diversity to "biodiversity" (Medina-Franco *et al.*, 2014; Koutsoukas *et al.*, 2014): diversity in terms of how the molecule interacts with the protein. Activity cliffs, structurally and chemically similar pairs of molecules that have a large difference in potency, indicate that chemical diversity may not be the best descriptor to identify molecules that interact with the target in sufficiently diverse ways. Activity cliffs arise from very similar molecules and consequently would not be selected by diversity methods. The molecules to be taken forward should be both active against the target and diverse in terms of how they interact with the target. Additionally, molecular diversity methods only based on the ligand can underestimate the diversity of two molecules in terms of the interactions of the molecules with the target. So far, the molecular diversity measures discussed have been ligand-centric, only measuring characteristics of the ligand, without any consideration of protein-ligand interactions.

LLOOMMPPAA (Bradley *et al.,* 2015) is a computational tool that suggests candidate compounds for follow-up as part of hit-to-lead optimisation using PLIFs: protein-ligand interaction fingerprints (Radifar *et al.,* 2013). The tool has been used as a

foundation for the CRANkS algorithm developed in this work. The algorithm takes fragment hits and enumerates a library of potential derivative compounds using a set of commonly used synthetic reactions (Cox *et al.*, 2016; Roughley *et al.*, 2011; Hartenfeller, *et al.*, 2011; Hartenfeller *et al.*, 2012). LLOOMMPPAA then makes use of 3D Matched Molecular Pairs to perform constrained conformer generation, to allow protein-ligand interactions to be calculated. PLIFs can then be generated from the protein-ligand interactions.

PLIFs are bit vectors where each bit represents whether an interaction is formed between a protein interaction point and the ligand: 1 if an interaction is present and 0 if the interaction is not, as shown in Figure 1.9. Once PLIFs have been generated for each candidate compound, diverse sets of compounds can be selected using the same dissimilarity algorithms as for ligand-centric fingerprints. Although this method includes protein-ligand interactions, and therefore should provide a better description of each candidate compound by including how it interacts with the target, it is subject to the same bias when compounds are selected from the diverse set for synthesis. Additionally, no knowledge is included about whether particular interactions are thought to be important for binding - all interactions are considered as equal. With the large increase in experimental data available, it is likely that a more informed selection of molecules for follow-up can be formed by considering interactions that have been observed before in the data.



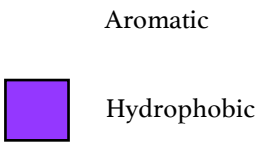



Figure 1.9 Representation of PLIFs used in LLOOMMPPAA (Bradley *et al.*, 2015). Each PLIF is a bit vector where each bit represents whether an interaction is present between the ligand and a protein residue.

## *1.5.2 Molecular Novelty*

Orthogonal to molecular diversity is molecular novelty. Rather than select a subset of molecules to maximise diversity that can then be further sub-selected for synthesis, introducing bias, molecules could be ranked based on how novel the molecule is. Medicinal chemists can then select from the ranking. There is no agreement on a clear definition for molecular novelty. The term is used in the literature to cover a number of descriptions including molecular dissimilarity, novelty in terms of the targets the molecule binds to, the shape of the molecule and the reaction used to synthesise it (Brown *et al.,* 2015) as well as the ring systems found within the molecule (Taylor *et al.,* 2014).

Molecular dissimilarity can be considered to be a measure of novelty: if molecule A is less similar to a set of molecules than molecule B, molecule A is more novel than molecule B. Molecular dissimilarity metrics have mostly been used as part of molecular diversity measurements. These often make use of fingerprints to encode the features of the molecule that are then compared using a dissimilarity or similarity calculation (see Section 1.3.1 and Section 1.5.1).

Molecular shape has also been used as a molecular descriptor to measure novelty. An analysis of the synthetic reactions used in medicinal chemistry was undertaken, and the effect on novelty was measured using chemical shape space (Brown *et al.*, 2015). Frequent use of the same synthetic reactions was found to have biased exploration to particular parts of chemical shape space showing how molecular novelty can be used to describe whether the molecule is synthesised in a novel way.

Novelty has also been described using the ring systems within a molecule. An analysis of the ring systems in drugs in the FDA Orange Book found that successful drugs consistently use the same ring systems and only 28% of approved drugs exhibit a ring that is different (Taylor *et al.*, 2014). Thus, molecular novelty could be described in terms of whether the molecule consists of one of the conserved ring systems or contains a new ring system not previously seen.

There are methods that currently calculate a novelty score including the Activity Atlas package from Cresset (Cresset, Forge). Candidate compounds are aligned in 3D and a grid is generated that covers the aligned compounds. A coefficient is then calculated at each grid point given the steric and electrostatic fields of the molecules. A novelty score for each molecule is calculated using the generated grid that should assess which regions of the candidate compounds have fully been explored, disregarding biological activity.

This novelty score, and the definitions of molecular novelty described above are all ligand-centric. In this work, the CRANkS algorithm defines novelty based on how novel the candidate compounds are compared to known binders to the target, in

terms of chemical structure, placement in the binding site and their interactions with the protein. The inclusion of protein-ligand interactions should allow new compounds that satisfy interactions previously identified to be formed by known binders to be prioritised, by making use of available experimental data.

## 1.6 Fragment-Based Drug Discovery

High-throughput screening (HTS) was considered to be revolutionary for hit detection when it was first introduced (Lipinski *et al.*, 2001). Despite the increase in spending on research and development to enable high-throughput screening, the problems associated with the method became clear when the subsequent increase in the number of approved drugs was less than expected. High-throughput screening yields a high number of false positives – if millions of molecules are screened even a small false positive rate results in a large number of false positive results (Bibette, 2012). Additionally, the attrition rates for leads developed from high-throughput methods are high due to the limited chemical space that can be realistically navigated by this method (Cook *et al.,* 2014).

Fragment-based screening aims to circumvent some of these issues by screening smaller molecules against the target, typically having a molecular weight of less than 300 Da. This allows a more efficient exploration of chemical space. There are approximately $10^{10}$ synthesisable fragment-sized compounds (Shoichet, 2013). By screening 1000 fragment-sized compounds a greater proportion of the chemical space can be searched than by screening 1 million molecules of the typical size used in HTS, approximately 23 – 35 heavy atoms (Shoichet, 2013). Additionally, the reduced activity of fragments against targets, but higher ligand efficiency due to the

reduced size of the compound, allow issues such as molecular obesity to be prevented more easily (Hann, 2011). The "Rule of Three" is typically applied to fragment libraries, which states that all fragments should have a molecular weight of less than 300 Da, a clogP of less than three, fewer than three hydrogen bond donor atoms, fewer than three hydrogen bond acceptor atoms, and fewer than three rotatable bonds (Congreve *et al.,* 2003).

The advances in structure-based drug discovery described in Section 1.3 have allowed an increase in the use of fragment-based drug design. Starting from a fragment hit there are three main strategies for developing a fragment hit into a lead compound, shown in Figure 1.10. The first is fragment linking. This involves joining together two of the fragment hits by a chemical linker. The second approach is fragment growing. One of the fragments is grown using knowledge of the protein binding site to extend the chemical structure to form further interactions. The third is fragment merging, which involves developing a follow-up compound capable of forming interactions formed by two of the fragment hits that are bound in close proximity to each other (Hung *et al.,* 2009; Kozakov *et al.,* 2015).

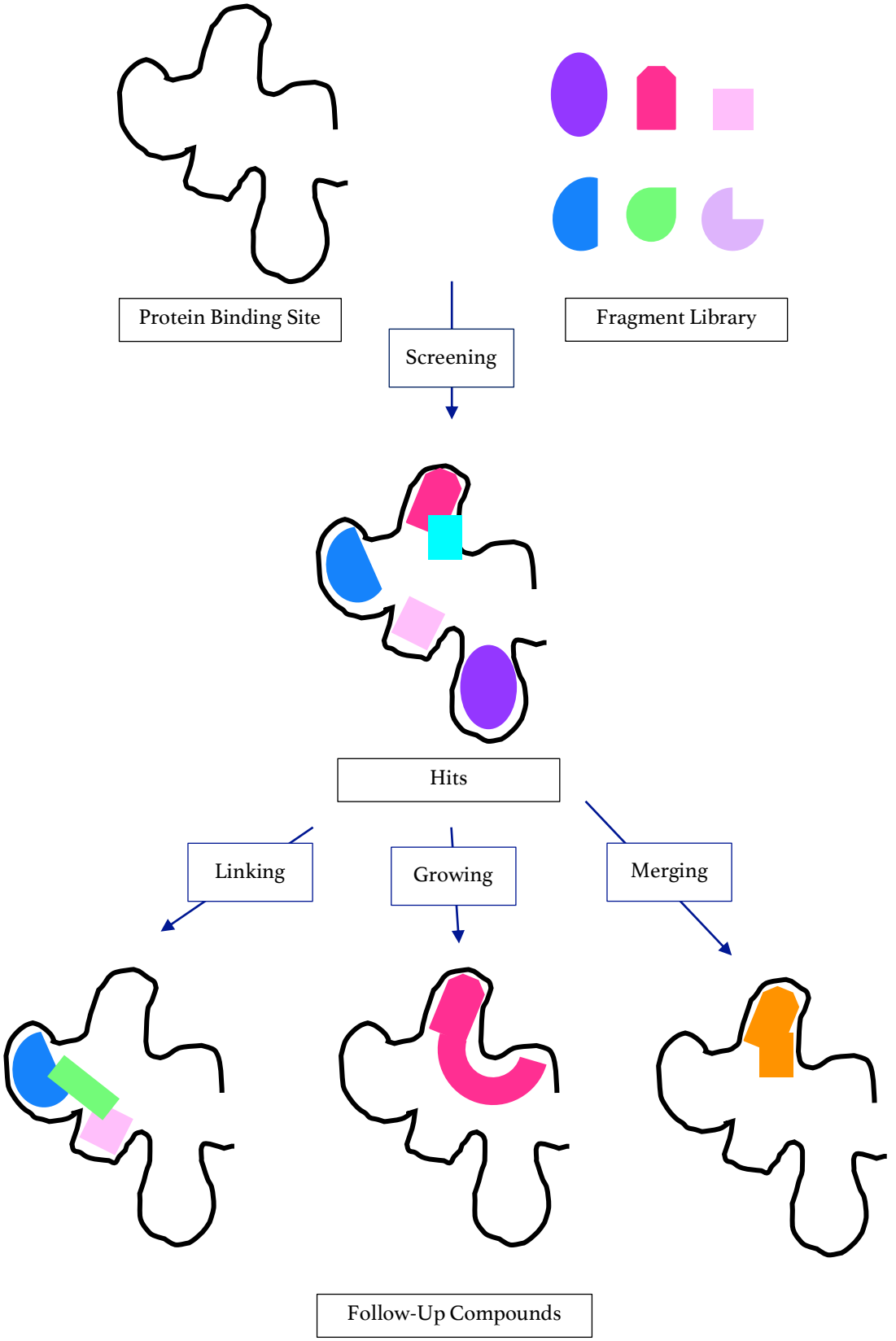



Figure 1.10 The process of developing fragment hits as part of hit-to-lead optimisation in fragment-based drug discovery. This can be by fragment linking, fragment growing or fragment merging.

Fragment screening is made more difficult due to the weak interactions between the protein and fragment. Typically, there will only be one or two directional interactions making a significant contribution to the binding affinity. This requires the assays used to determine whether a fragment is active against a given target to be much more sensitive than for high-throughput screening. Examples of biophysical methods used for fragment screening include surface plasmon resonance, fluorescence-based thermal shift, $^{1}$H NMR spectroscopy and X-ray crystallography (Christopher *et al.,* 2013; Silvestre *et al.,* 2013; Erlanson *et al.,* 2016).

X-ray crystallography has only recently been used as a primary screening method and traditionally was used as a secondary screen or for target validation (Erlanson *et al.,* 2016). However, at the *i04*-1 beamline at Diamond Light Source, a process has been compiled that allows 1000 compounds to be screened using X-ray crystallography within one week by streamlining the harvesting and soaking technique and developing methods for fast data analysis. One compound can be soaked per crystal, allowing 100 to 1000 times higher concentrations of the compound with respect to the target than for traditional solution-based assays. This allows increased sensitivity for the weak-binding fragment hits. The recent development of the PanDDA method has allowed for higher confidence modelling of the weakly bound hits obtained by fragment screening (Pearce *et al*., 2017).

This screening yields a large amount of data about the protein structures with bound hits. There have been a number of successful uses of fragment-based drug design that used fragment binding models for hit-to-lead optimisation. However, there is currently no method that can make use of this structural data, with multiple

instances of hits binding to the same target, to guide hit-to-lead optimisation in a systematic method. This sets the precedence for this thesis – I describe the development of algorithms that make use of protein-ligand structures for a given target to guide hit-to-lead optimisation.

## 1.7 Project Aims

With high attrition rates, and the increasing costs of each stage of drug development, there is an emphasis on making the right decisions in the earliest stages of drug discovery. Hit-to-lead development and lead optimisation are currently dominated by subjective-decision making based on synthetic intuition, experience and existing bias. Therefore, there is an emphasis on developing methods that will prioritise which compounds to make next as part of the hit-to-lead process that will avoid this subjective decision-making.

In Chapter 2 I describe an algorithm I have developed called CRANkS (Compound Ranking for Active Novel Scaffolds) to prioritise candidate compounds based on how novel the compounds are compared to known inhibitors of the target of interest. The standard way to calculate novelty is molecule-centric. In this work, I approach the problem from an interaction-centric view and aim to discover both chemical novelty and potential novelty of interaction. A prospective compound is compared to known binders of the target in terms of its predicted interactions with the protein, placement in the binding site, and chemical structure. Scalar grids are generated to capture the spatial distribution of the chemistry of the known ligands within the

binding site, and vector grids are generated around the protein active site to describe the protein-ligand interactions. Candidate compounds to be scored are placed in the binding site using 3D Matched Molecular Pairs (Bradley *et al.*, 2014) and constrained conformer generation. Three novelty scores are calculated: the Element Score, the Pharmacophore Score and the Interaction Score. I tested the CRANkS algorithm on active and inactive compounds for HIV-1 protease and discuss possible improvements of the method.

The preliminary results were promising, and this prompted modifications to the algorithm before a larger testing of the CRANkS algorithm. In Chapter 3 I discuss the testing I have performed of the CRANkS algorithm on a number of benchmark datasets for different targets. By using the calculated novelty scores, I hypothesised that the CRANkS algorithm will prioritise both compounds that are active, but also compounds that are novel scaffolds. Scaffold-hopping is of particular importance in drug discovery for a better exploration of the activity space, in order to find the most amenable lead compound to take forward, in terms of bioactivity, but also physiochemical properties such as toxicity. The ability of the CRANkS algorithm to prioritise active compounds was tested by calculating metrics to describe the ranking of compounds for benchmark datasets from the Database of Useful Decoys Enhanced (Mysinger *et al.,* 2012). The values for the metrics were compared to those from popular commercial docking tools and other virtual screening methods. The success of the CRANkS algorithm was found to be target-dependent, but overall performed as well as commercial docking methods. Metrics were also calculated to assess the scaffold-hopping potential and the algorithm was found to outperform commonly used similarity methods. Finally using multi-objective optimisation, the

algorithm was found to select subsets of follow-up compounds from these datasets enriched for novel active scaffolds.

In Chapters 2 and 3 I have shown the potential of the CRANkS algorithm to separate actives from inactives. To determine whether this signal could aid the docking process I have developed "active-guided docking" by combining the grids developed as part of CRANkS with the AutoGrid part of AutoDock – AutoCRANkS - and this is described in Chapter 4. This combination uses structural data of protein-ligand complexes to modify the scoring function and conformer generation to be target-specific in line with the "one size does not fit all" conclusions in relation to docking scoring functions (Ross *et al.,* 2013). Conformer generation using AutoDock also circumvents issues with the conformer generation part of the CRANkS. I tested different weightings of combining the CRANkS grids with grids used by AutoDock for a number of targets and investigated different combinations of modifying the scoring function, the conformer generation or both. Overall AutoCRANkS was found to have a target-dependent effect on the performance of the docking algorithm but for some targets performance was greatly improved. Further testing on a large number of targets is required now the method has shown promising results, and specific parameters for combining the CRANkS grid with AutoDock have been determined.

Chapter 5 describes testing the CRANkS algorithm in a prospective setting at Diamond Light Source, making use of the increased experimental data now available at the start of the hit-to-lead process. This highlighted potential issues in using the

algorithm in a drug discovery environment particularly in terms of the conformer generation.

In this thesis I have developed the CRANkS algorithm and AutoCRANkS to make use of the increase in experimental data at the initial stages of drug discovery to aid hit-to-lead optimisation by selecting sets of compounds enriched with active scaffolds for synthesis. This will aid rationally driven selection of compounds. Using multi-objective optimisation and a data-driven approach, a diverse set of activity-enriched compounds can be taken forward for synthesis, making use of the knowledge from the hit identification stage, protein-ligand interactions as well as predicted potency.

# Chapter 2

# CRANkS Algorithm

## 2.1 Introduction

I have developed a computational method called CRANkS (Compound Ranking for Active Novel Scaffolds) to score compounds based on how novel the compound is compared to known binders of the target of interest. An overview of the algorithm developed is shown in Figure 2.1. Structural data of protein-ligand complexes are aligned and used to generate three CRANkS grids: the element grid, pharmacophore grid and interaction grid. These grids describe the binding pocket in terms of the element types of atoms of the ligand, its pharmacophoric features and the protein-ligand interactions made by each active ligand.

Candidate compounds to be scored are placed in the binding site using constrained conformer generation: coordinates are assigned to part of the molecule by finding 3D Matched Molecular Pairs between the candidate molecule and the ligands in the structural data (Cumming *et al.*, 2013; Ehmki and Rarey, 2018). Conformations are then generated for the rest of the molecule. Each conformer is scored based on how it overlaps with each of the grids generating three novelty scores per conformer: the Element Score, the Pharmacophore Score and the Interaction Score. These scores can then be used to rank compounds by novelty and prioritise either scaffold-hopping or compounds that form novel protein-ligand interactions. Each step of the

algorithm is stored in a database allowing the results to be viewed by a web-based viewer. The viewer has been developed to interpret scores by placing conformers within representations of each of the grids. Each step of the algorithm will now be discussed in turn. The CRANkS algorithm was developed primarily using Django (Django, 2015) with Python (van Rossum, 1995) and RDKit (Landrum, 2015).

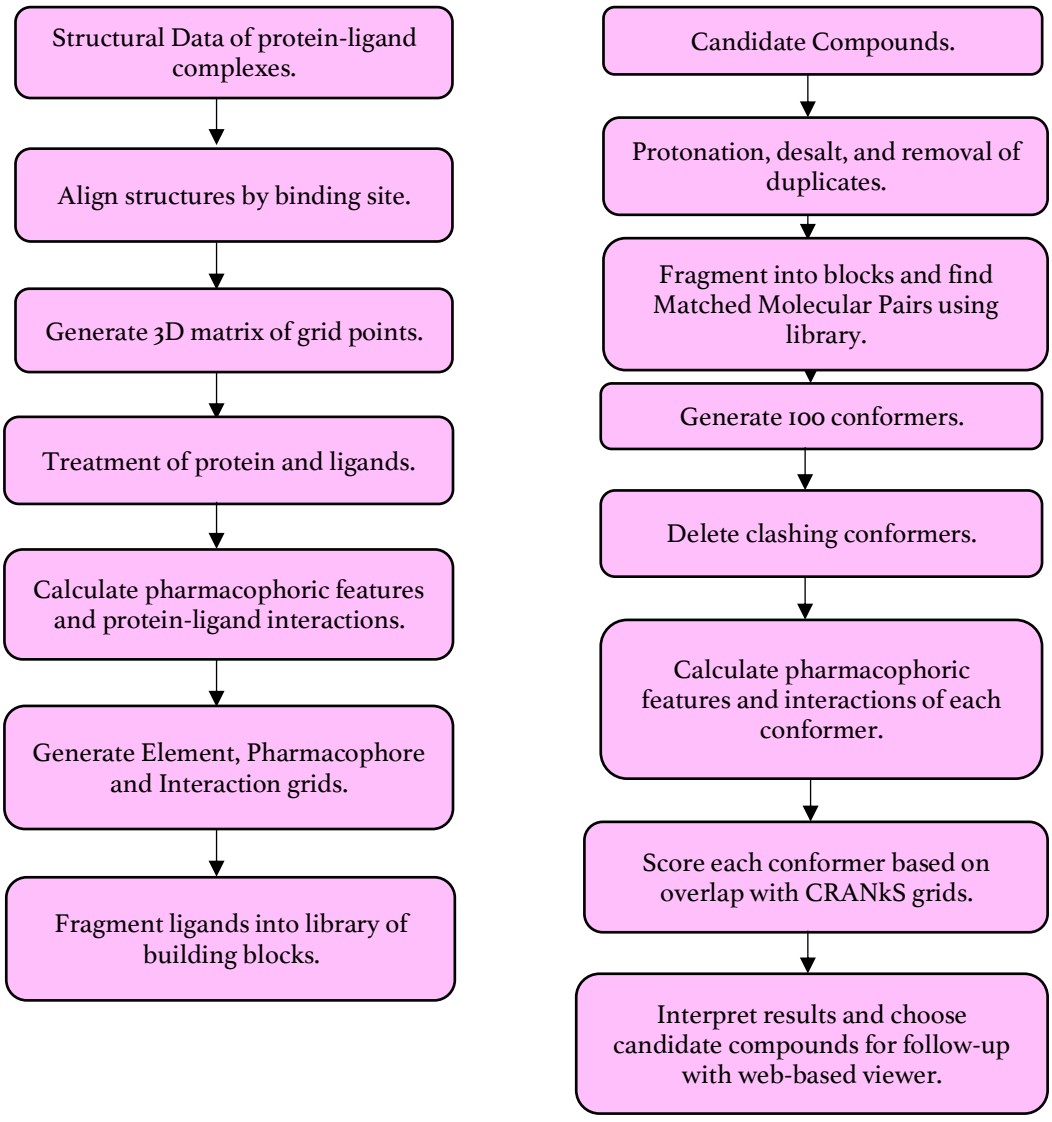



Figure 2.1. Depiction of the algorithm developed to score candidate compounds based on novelty. The chart on the left shows the treatment of structural data of known protein-ligand complexes. The right shows the processing of candidate compounds to be scored. Each of these steps are discussed below.

## 2.2 Algorithm

### *2.2.1 Grid Generation*

Grids have frequently been used as part of *in silico* drug design. The original use of a grid-based system for drug discovery can be attributed to Peter Goodford and the GRID program (Goodford, 1985, Wade *et al*., 1993). This was designed to determine binding sites on proteins that are energetically favourable by calculating the interaction of a probe with a protein target at sequential grid points using an energy function. The results are displayed as an isoenergetic 3D contour and the software has been used to detect binding sites for use in molecular design. Grids have since been used regularly to represent protein-ligand systems as part of docking or binding site detection. Grid points can be used to store energy potentials that represent physiocochemical properties or features, scores from scoring functions or energy values calculated using force fields. Grids also lend themselves to trilinear interpolation, which can be used to calculate interaction energies quickly - this is used by AutoDock 4 for example (Morris *et al*., 1998).

In this case a pre-computed grid-based system offers computational speed compared to considering continuous space or running expensive molecular mechanics simulations. The grid provides a sufficient level of detail to describe the system, but also a level of simplification to capture trends and patterns within the structural data. The coarseness of the grid should allow for some error in coordinates and alignment and also help account for the fact that static 3D-coordinates are being used to

describe a dynamic system. Conversely, the grid needs to be fine enough to distinguish between different atoms in the ligands and to accurately describe the 3D distribution of groups within the binding pocket. The length of a carbon-carbon single bond is approximately 1.54 Å. Therefore, a grid spacing of 1.0 Å was chosen as this should be able to assign two atoms bonded to each other to different grid points whilst offering a sufficient level of coarseness. This should also describe the binding site using a small enough number of grid points to allow the algorithm to use less memory and to be quick to run.

Before grid points are generated, the structural data that the grid describes must be aligned. The protein-ligand complexes are aligned using the *align* function in PyMOL (Schrödinger, LLC). Each protein-ligand complex to be used in the grid is loaded into PyMOL (Schrödinger, LLC). The user can then align the protein-ligand complexes manually: in the case of HIV-1 protease the complexes are aligned by the binding site using residues 20 to 30 around catalytic residue Asp-25. The binding site must then be defined by the user - the user must provide the maximum and minimum $x$, $y$ and $z$ coordinates of a cube that covers the binding pocket. A 3D matrix of grid points is then generated which covers the binding site using these coordinates. This is generated using *NumPy* within Python (Hugunin, 1995). An example of generated grid points is shown in Figure 2.2 for a set of HIV-1 protease protein-ligand complexes.

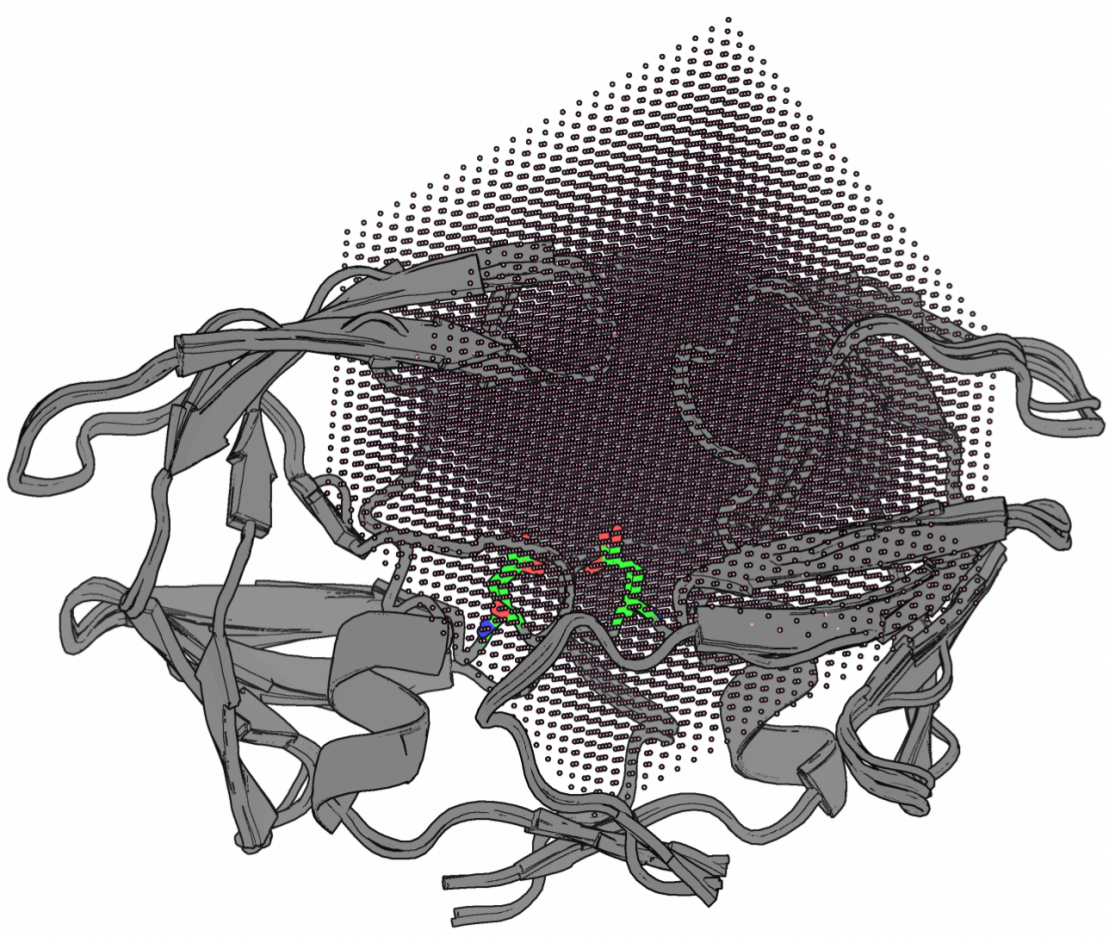


Figure 2.2. 3D matrix of grid points represented by dots generated to describe the binding site of HIV-1 protease. Each grid point is represented by a sphere and the spacing between grid points is 1 Å. The catalytic residues shown in green define the binding site and the grid is shown to cover this.

### *2.2.2 Treatment of Protein and Ligand and Calculation of Features*

To be able to assign pharmacophoric features and calculate protein-ligand interactions correctly, both the protein and ligand first need to be protonated at the correct pH. Protonation can affect the behaviour of functional groups and charges and thus the ability to form interactions, such as salt bridges. For example, a protonated carboxylic acid group will exhibit different properties if deprotonated.

The deprotonated group will act as a hydrogen bond acceptor and form different interactions to the protonated group, which will act as a hydrogen bond donor. Consequently, care must be taken to assign protonation states and to ensure both ligand and protein are protonated to the same pH. Both the protein and ligand are protonated to a pH of 7.4 - approximately the physiological pH of blood in the body which is discussed below.

#### *2.2.2.1 Protein*

The protein is protonated using the command line version of PDB2PQR (Dolinsky 2004, Dolinsky 2007). This uses the PROPKA algorithm (Li *et al.,* 2005) to estimate titration states and add hydrogens to proteins for a given pH, outputting a PQR file. This is then converted back to a PDB file using Open Babel (O'Boyle, 2011), allowing each protein chain to be read from the file and converted to an RDKit molecule.

Pharmacophoric features are then calculated for each protein chain using the RDKit feature factory. A subset of protein-specific functional groups was used to create the factory - this subset is a modified version of that used by Bradley *et al.*, (Bradley *et al.,* 2015), which now includes carbonyl groups as hydrogen-bond acceptors. The SMARTS expressions used to define functional groups that are classified as a pharmacophoric feature are shown in Figure 2.3. A diagram has been generated using SmartsViewer (Schomburg *et al.,* 2010) to accompany each SMARTS string to explain the string. The functional groups fall into six family categories that broadly describe their behaviour: acceptor, donor, negatively ionisable, positively ionisable, aromatic and hydrophobe. Within these families the SMARTS are divided into further groups.

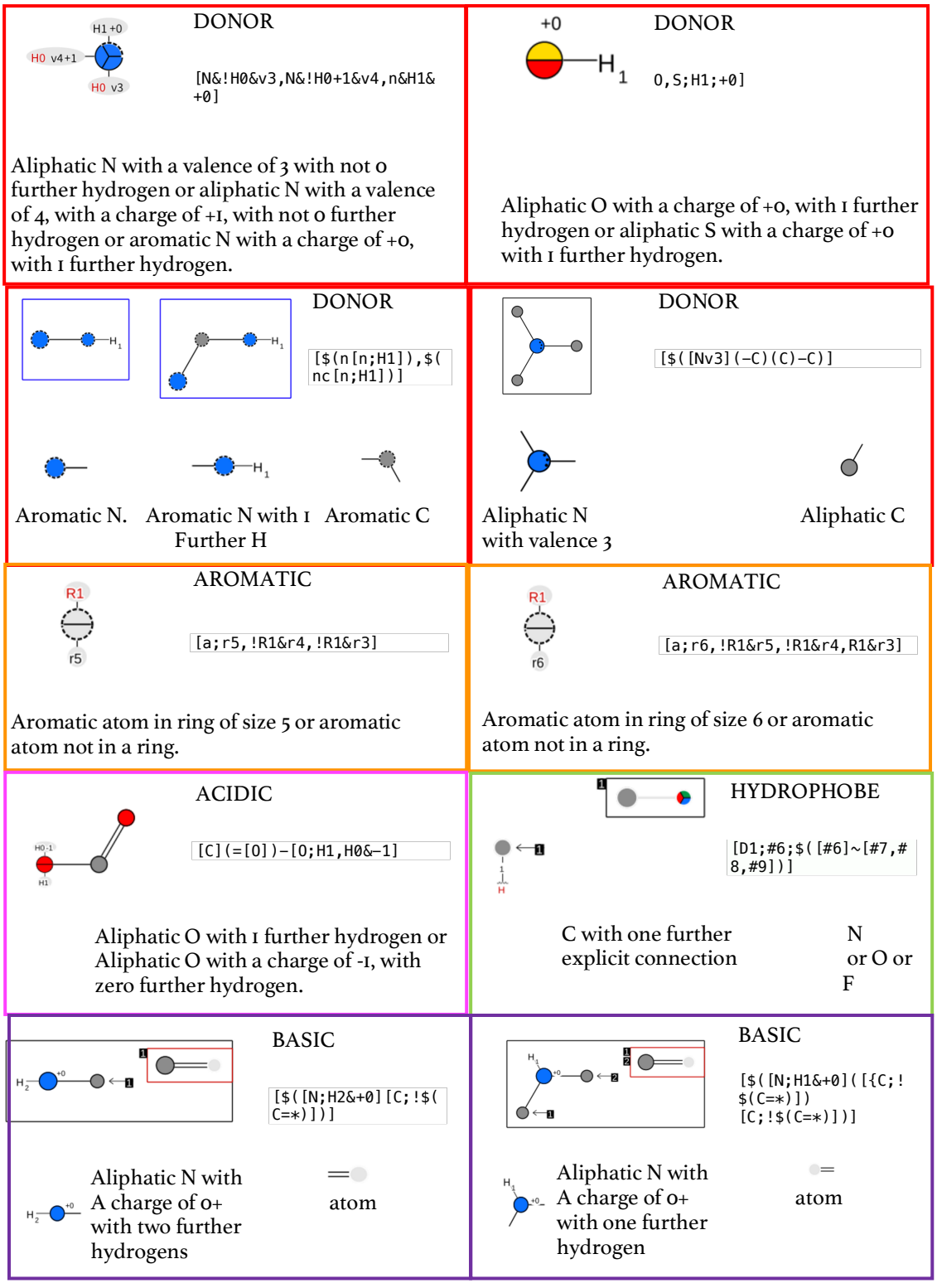



Figure 2.3. *(continued on next page)*

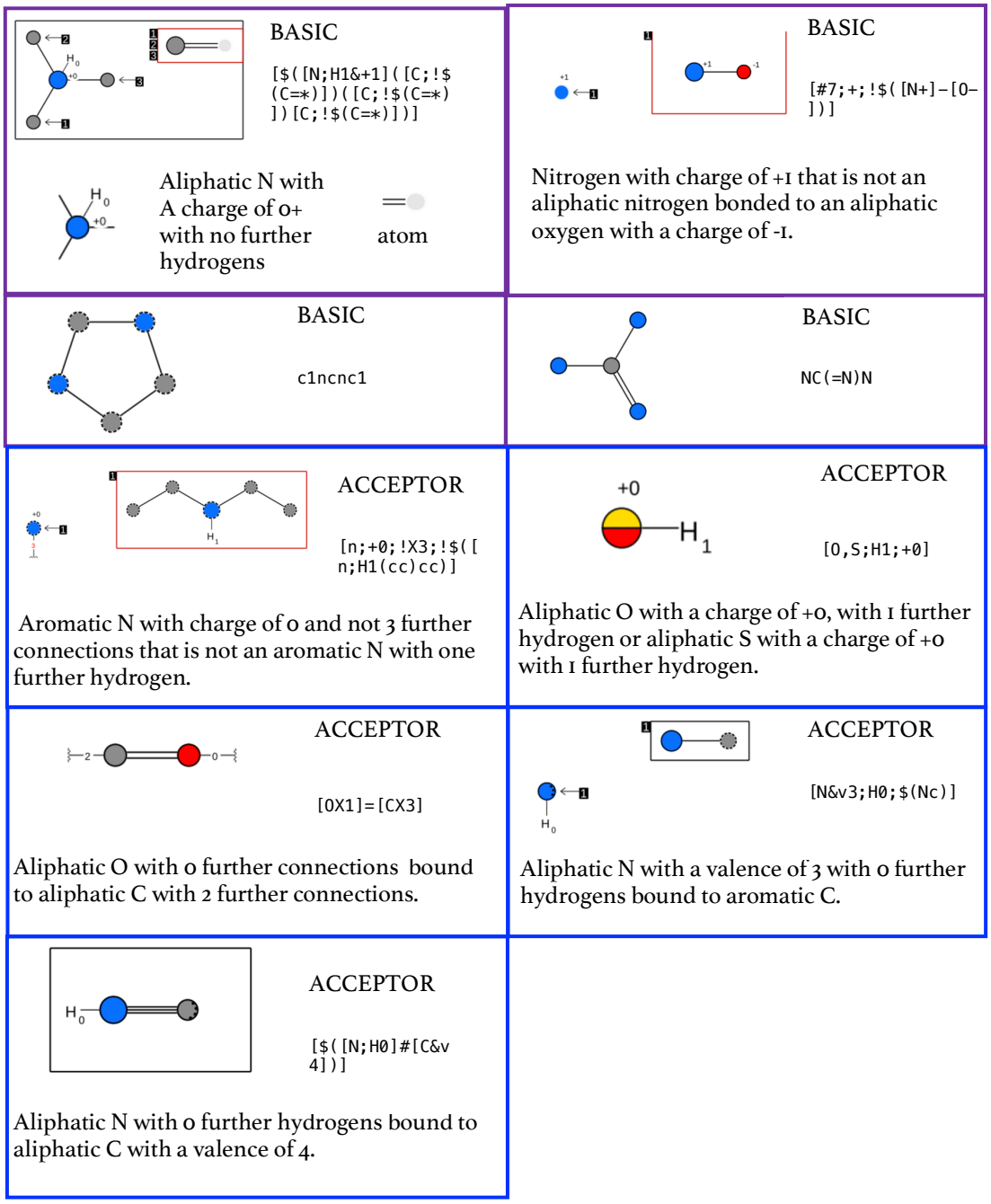



Figure 2.3. (continued from previous page). Table of SMARTS expressions detailing the pharmacophoric features calculated for a protein. The visualization of the SMARTS is adapted from SMARTSViewer (https://smartsview.zbh.uni-hamburg.de).

Notably, in Figure 2.3, the SMARTS *"[O,S;H1;+0]"* are used to define both single atom acceptors and donors. This reflects that the hydroxyl group can both donate its H group to form hydrogen bonds, but also accept protons to form hydrogen bonds due to the two oxygen lone pairs. Thus, it acts as both an acceptor and a donor. This illustrates the flexibility of this method to capture the nuances of the behaviour of chemical groups. This method of assigning pharmacophoric features also demonstrates its usability as the feature is assigned to a number of atoms within the group and also given a central coordinate. In the case of carboxylic acids, which due its two resonance states could be protonated or charged on either oxygen atom, both oxygen atoms are assigned the pharmacophoric feature. Consequently, interactions can be calculated from both oxygens so that the calculations are less dependent on the protonation of the protein.

#### *2.2.2.2 Ligand*

The ligand is first extracted from the PDB file containing the protein-ligand complex and written to a separate PDB file containing only the ligand. If there are multiple occupancies, each conformer is saved to a separate file. This is then converted to an SDF file using Open Babel (O'Boyle *et al.*, 2011), allowing the ligand to be protonated using *cxcalc* from the command line version of Calculator Plugins (ChemAxon, 2016). The protonated ligand is output as an SDF file which is then converted to an RDKit molecule. The SMILES string for the ligand, taken from the Protein Data Bank (Berman *et al.,* 2000), is required to assign the correct bond order and connectivity to the ligand. Pharmacophoric features are then calculated using the inbuilt feature

factory of RDKit (Landrum, 2015) - as the ligands are small molecules, a reduced set of definitions of features was not required for computational efficiency.

### *2.2.2.3 Protein-Ligand Interactions*

Once the pharmacophoric features on both the protein and ligand have been calculated, protein-ligand interactions are calculated between the features. The interaction definitions used are from the work of Bradley *et al.*, (Bradley *et al.,* 2015) which in turn are defined using the Protein- Ligand Interaction Profiler (Salentin *et al.,* 2015). Four interactions are calculated: hydrogen bonds (acceptor-donor), acid-base, hydrophobic and π-π interactions. The definitions of these interactions are shown in Table 2.1. In the work of Bradley *et al.*, (Bradley *et al.*, 2015) a broad definition was developed to help account for the flexibility in a dynamic protein-ligand complex, and the likely error in conformations, that are represented by static coordinates. The ligands and protein chains, as well as the atoms, pharmacophoric features and interactions, are stored in the CRANkS database as discussed in Section 2.9.

| Interaction | Pharmacophore 1 | Pharmacophore 2 | Distance / Å | Distance (Broad) / Å | Angle / ° | Angle (Broad) / ° |
|---|---|---|---|---|---|---|
| Hydrogen-Bond | SingleAtomAcceptor | SingleAtomDonor | 3.5 | 4.5 | 120:360 | 90:360 |
| Acid-Base | AcidicGroup | BasicGroup | 4.5 | 5.5 | — | — |
| Hydrophobic | Hydrophobe | Hydrophobe | 4.0 | 5.0 | — | — |
| Aromatic | Arom5/Arom6 | Arom5/Arom6 | 4.5 | 5.5 | — | — |

Table 2.1. Interactions definitions used by the algorithm from the Protein-Ligand Interaction Profiler and LLOOMMPPAA. Arom5/Arom6 refer to five or six-membered aromatic rings.

### *2.2.3 Calculation of Element, Pharmacophore and Interaction Grids*

Once the structural data is stored in the CRANkS database, CRANkS grids are generated to describe the data. Compounds can then be scored on how the compound overlaps with the grids. Three grids are generated: the element grid, pharmacophore grid and interaction grid. These describe the location and element type of the atoms of the bound ligands, the pharmacophoric features of these ligand, and their protein-ligand interactions.

#### *2.2.3.1 Element Grid*

To generate the CRANkS element grid, the algorithm sequentially takes each atom of each ligand. The nearest grid point to the atom is found. Each grid point holds a set of counts for each type of element found at that grid point. The crystallographic occupancy of the atom is added to the corresponding count at the nearest grid point given the element of the atom. For example, if I take a carbon atom of occupancy 1.0, 1 is added to the carbon atom count at that grid point. If at the end of the grid generation the grid point was the closest grid point to 12 carbons all with occupancy 1.0, the carbon atom count at that grid point would be 12. Counts are stored for the following elements: carbon, oxygen, nitrogen, sulphur, and all elements found in the set of known ligands.

### *2.2.3.2 Pharmacophore Grid*

The pharmacophore grid is generated by looping through all the pharmacophoric features of each ligand. Each grid point holds a set of counts for each type of pharmacophoric feature. Given a pharmacophoric feature, the nearest grid point is found. The crystallographic occupancy of the ligand atom is added to the count for the given type of pharmacophoric feature. For example, for a ligand atom with occupancy 1.0 with an acceptor pharmacophoric feature, 1.0 is added to the count of acceptors at the nearest grid point to the feature.

The pharmacophoric features calculated using RDKit are classified into broader family groups and then are further divided into more detailed types. Currently, the pharmacophore grid uses the broad pharmacophore family groups and counts are stored for acceptor, donor, hydrophobe, lumped hydrophobe, aromatic and ionizable features.

### *2.2.3.3 Interaction Grid*

The interaction grid is a vector grid and is calculated in a different way to the other two scalar grids. Interactions are calculated and stored as a vector between two grid points. Therefore, rather than counts stored at individual grid points, the interaction grid is represented by interaction counters. Interaction counters hold the number of a given type of interaction between two grid points. Each protein-ligand interaction is taken in turn, and the start and end grid points are identified. The database is searched for an interaction counter between the two grid points for the type of the

interaction. If there is an interaction counter 1 is added to the count. Otherwise an interaction counter is created with a count of 1.

### *2.2.4 Treatment of Candidate Compounds*

Once the grids have been created, the candidate compounds to be scored are read into the database. However, the molecules are first subject to a pre-processing pipeline shown in Figure 2.4. The compounds are converted into RDKit molecules from an SDF file. First the molecules are checked against the ligands used to create the grid and any duplicates are removed. Next the molecules are desalted using the RDKit Salt Remover function (Landrum, 2015). Molecules with ambiguous chiral centres are removed to avoid issues with stereoisomers. The molecules are also subjected to a PAINS filter. PAINS: Pan Assay Interference CompoundS (Baell *et al.*, 2010), are molecules likely to interfere with assay results, by either affecting the way the assay measurements are taken, or by being highly reactive with protein targets. The filter is achieved by using RDKit to find molecules that match a list of SMARTS patterns that refer to PAINS and remove the matching molecules. The remaining molecules are then written back to an SDF file and protonated to pH 7.4 using *cxcalc* (ChemAxon, 2016). Finally, the compounds are stored in the database (see Section 2.9) in mol block format.

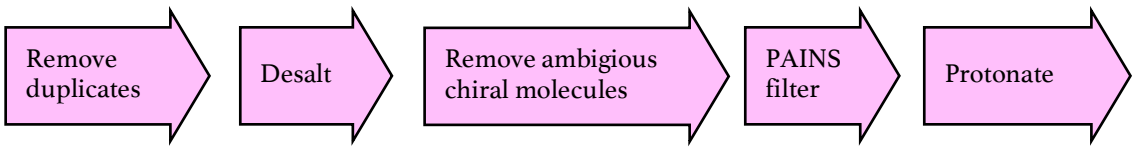



Figure 2.4. Flowchart describing the treatment of candidate compounds before the compounds are saved to the database.

### *2.2.5 Conformer Generation*

Candidate compounds are placed into the binding site to be scored. 3D conformations need to be generated for each compound. The method developed here uses a constrained conformer generation, making use of the 3D Matched Molecular Pairs method used by A. R. Bradley *et al.,* (Bradley *et al.,* 2015). By using constrained conformer generation, I hypothesise that more accurate conformations can be produced, by making use of the structural data available to copy coordinates from the matched atoms in the X-ray crystal structure.

#### *2.2.5.1 3D Matched Molecular Pairs*

To perform the constrained conformer generation, first the part of the molecule that will be constrained must be identified. In this work, this is achieved using 3D Matched Molecular Pairs (Cumming *et al.,* 2013; Ehmki and Rarey, 2018). First, the ligands from the structural data used to construct the grids are fragmented. The SMARTS expression used to select atoms that are connected by bonds to be broken is *"[#6+0;!$(*=,#[!#6])]!@!=!#[*]"* which is depicted in Figure 2.5. Bonds are broken that are single bonds and are not within a ring. The atom in question must be connected to a carbon atom. That carbon atom cannot be connected to another carbon atom and cannot form double or triple bonds. An example of the result of the fragmentation is shown in Figure 2.6. The building blocks the molecule breaks into are stored in the database.

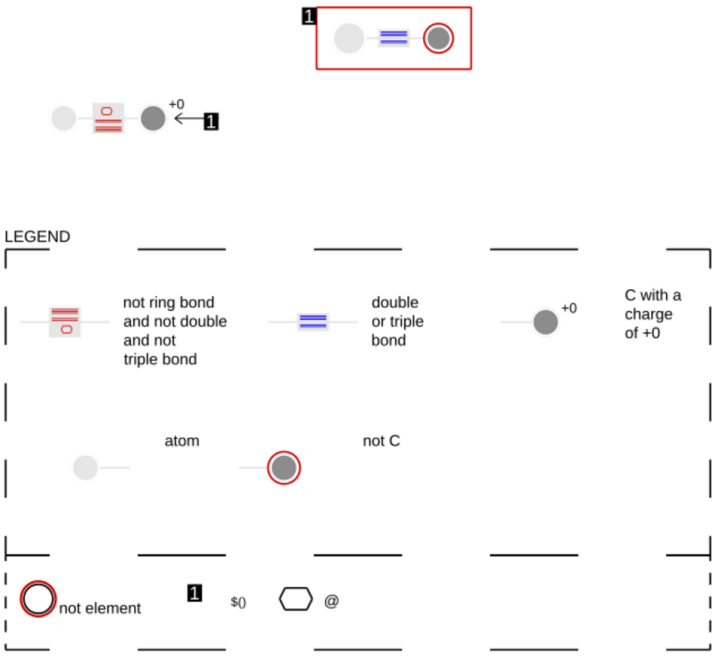



Figure 2.5. SmartsViewer depiction to describe the SMARTS "[#6+0;!$(*=,#[!#6])]!@!=!#[*]" used to identify atoms that form bonds to be broken to create building blocks for Matched Molecular Pairs analysis. The atom must be bound to a carbon with a single bond that is not part of a ring.

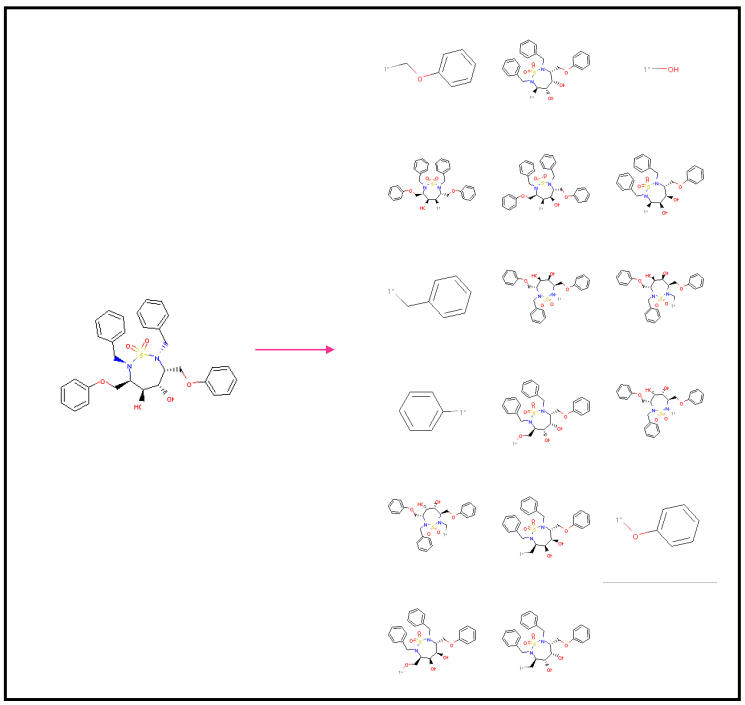


Figure 2.6. Ligand AHF from PDB structure 1g35 and the resulting building blocks formed after the ligand is fragmented using the SMARTS expression in Figure 2.5.

Next, each candidate compound is taken in turn and fragmented according to the same SMARTS expression. Each resulting block is then searched for a match in the library of building blocks. In the case that multiple matches are found, the match with the largest number of heavy atoms is taken forward. The matching ligand and candidate compound are known as a Matched Molecular Pair.

#### *2.2.5.2 Conformer Generation*

Once the 3D Matched Molecular Pair of the candidate compound and ligand is found the conformer can be generated using the *Constrained Embed* function of RDKit (Landrum, 2015). The building block found to match between the ligand and compound is used as a template to assign coordinates to the matching part of the candidate compound. A conformer is then generated for the rest of the molecule. The function achieves this by generating conformers using a distance geometry method: this involves calculating a distance bounds matrix using rules and the connection table of the molecule. A random distance matrix is generated which satisfies these bounds. In this case, the distance bounds matrix includes the atom-atom distances in the template i.e. distances from the ligand identified to be a Matched Molecular Pair with the candidate compound. The conformer is then optimised using the UFF force field before the conformer is aligned to the template. A depiction of the conformer generation is shown in Figure 2.7. One hundred conformers are generated for each candidate compound. Any conformer with an atom that occupies the same grid point as a protein atom is then deleted. This avoids any clashes with the protein.

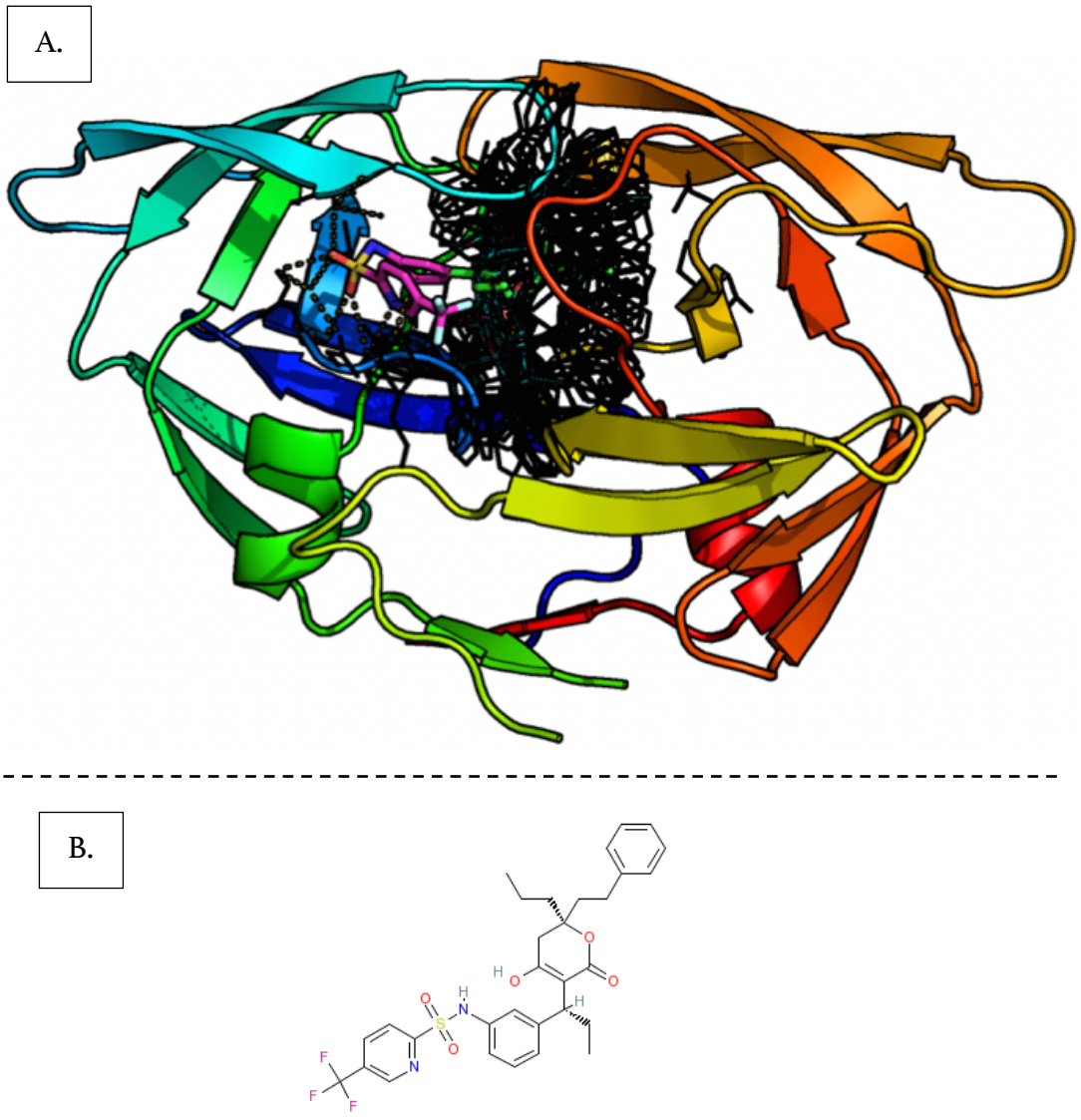



Figure 2.7. Conformers of a candidate compound within the HIV-protease active site is shown in A. In this case the candidate compound is ligand TPV from PDB structure 1d4y. The structure was tested within the CRANkS algorithm by splitting the ligand into blocks using the SMARTS expression discussed in Section 2.2.5.1. The ligand was then "redocked" by constraining with one of the available blocks to test the algorithm. The conserved matching block found is shown in pink. The 100 conformers generated for the rest of the molecules are shown in black, and the crystallographic structure of the rest of the compound is shown in green. Note that there are some steric clashes which have not been eliminated yet. B shows the 2D structure of ligand TPV.

## *2.2.6 Calculation of Scores*

### *2.2.6.1 Element Score*

The Element Score of a conformer is calculated by sequentially taking each atom of the conformer. The nearest grid point to the atom is found and the count at that point given the type of atom is taken, $N$. The maximum count for that element at any point in the grid is also taken $N_{max}$. The Element Score for the atom ($s_E$) is then calculated using Equation 2.1:

$$s_E = 1 - \frac{N}{N_{max}} \tag{2.1}$$

The Element Score for the whole conformer, $S_E$, is then the average of the Element Score of all the atoms in the conformer, where $n_c$ is the number of atoms in the conformer (Equation 2.2):

$$S_E = \frac{\sum_1^{n_c} s_E}{n_c} \tag{2.2}$$

Finally, the Element Score of the candidate compound is the average of the Element Score over all the conformers.

### *2.2.6.2 Pharmacophore Score*

The Pharmacophore Score is calculated in a similar way to the Element Score, as both grids used for scoring are scalar grids. For each conformer of a candidate compound the pharmacophoric features are calculated using the same method as for the ligands - the in-built feature factory of RDKit. For each pharmacophoric feature the nearest grid point is found. The count of that type of feature at that grid point is taken, $N$. The maximum count of that type of pharmacophoric feature at any grid point is also taken $N_{max}$. The Pharmacophore Score for the pharmacophoric feature, $s_P$ is then calculated using Equation 2.3.

$$s_P = 1 - \frac{N}{N_{max}} \tag{2.3}$$

The Pharmacophore Score for the whole conformer, $S_P$, is then the average of the Pharmacophore Scores for all pharmacophoric features in the conformer where $n_c$ is the number of features, as shown in Equation 2.4.

$$S_P = \frac{\sum_1^{n_c} s_P}{n_c} \tag{2.4}$$

The Pharmacophore Score for the candidate compound is then the average of the Pharmacophore Scores of the conformers.

### *2.2.6.3 Interaction Score*

The calculation of the Interaction Score involves vectors and a vector field rather than counts at single grid points. For each conformer, each protein-ligand interaction, *a*, is taken in turn. The interactions are calculated as described in Section 2.2.2.3. The interaction is defined as occurring between two grid points: the starting grid point referring to the location of a pharmacophoric feature on the small molecule, and the ending grid point, referring to the location of a pharmacophoric feature on the protein. All the interaction counters with the same type of interaction and starting point as *a* are found, *b,* as shown in Figure 2.8 with interaction counters $b_1$ and $b_2$. The Interaction Score, $S_I$, for the conformer is then calculated using Equation 2.5 where $n_i$ is the number of interactions the conformer is calculated to form.

$$S_I = 1 - \sum_1^{n_i} \hat{a}.\hat{b} \times \frac{N_b}{N_{max}} \tag{2.5}$$

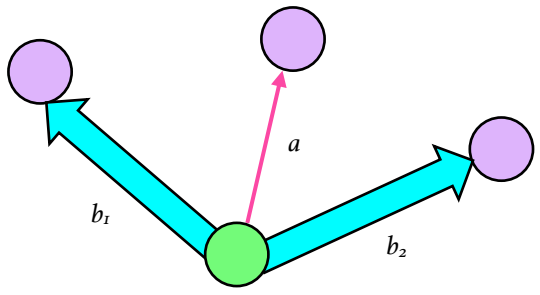



Figure 2.8 Protein-ligand interaction of a candidate compound, *a*, is shown in pink. It originates from the starting grid point in green, to a grid point on the protein in purple. Interaction counters $b_1$ and $b_2$ originate from the same starting grid point but end at different grid points. These are used to calculate the Interaction Score using Equation 2.5.

The count $N_b$, of each interaction counter, is found. This is then normalised by the maximum count for that type of interaction at any interaction counter $N_{max}$. This is then multiplied by the dot product between the unit vectors with the same direction as interaction counter *b* and interaction *a*. This allows the contribution of the interaction counter to be weighted by how similar the direction of the interaction counter is to the direction of the interaction. The Interaction Score for each individual interaction is then the average of this value over the interaction counters with the same starting grid point and type. The Interaction Score for the conformer, $S_I$, is then the average of the score for all protein-ligand interactions formed by the conformer. Finally, the overall Interaction Score for the candidate compound is the average of the Interaction Score for each conformer. The Interaction Score is therefore bound by zero and one with a high Interaction Score indicating high novelty and a low Interaction Score indicating low novelty.

#### *2.2.6.4 Interpretation of CRANkS Scores*

The three scores can be used to make decisions on which compounds to prioritise for synthesis and can be used to rank the candidate compounds. I hypothesise that the scores can be used to prioritise candidate compounds in two different ways as shown in Figure 2.9.

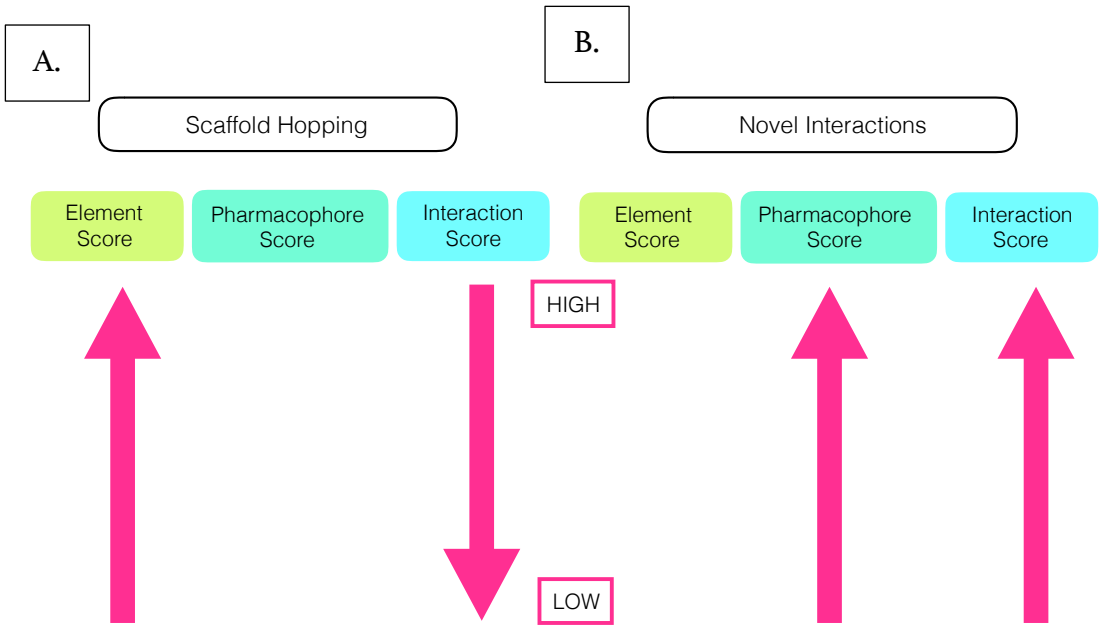



Figure 2.9. Explanation of how the CRANkS novelty scores can be interpreted. Scaffold hopping can be prioritised by taking forward molecules with a high Element Score and low Interaction Score (A). Novel interactions can be prioritised by taking forward molecules with a high interaction and Pharmacophore Score (B). The novelty scores are bound between zero and one with a high score indicating high novelty.

Scaffold-hopping (Figure 2.9 A) can be prioritised by selecting compounds with a high Element Score but a low Interaction Score. In this way, compounds that are less novel in terms of their protein-ligand interactions are prioritised. These compounds should therefore form interactions more similar to those previously seen in the grids, and therefore interactions that are likely to be important for binding. In addition, a high Element Score prioritises compound that are novel in terms of the chemical composition of the molecule and placement in the binding site.

Candidate compounds that form novel protein-ligand interactions can be prioritised by selecting compounds with a high Interaction Score and Pharmacophore Score (Figure 2.9 B). Compounds with a high Interaction Score should form interactions

that are novel compared to the protein-ligand interactions used to form the grid. A high Pharmacophore Score should indicate that the candidate compound possesses novel pharmacophoric features in terms of the placement of the ligand features in the binding site.

### *2.2.7 Web-based Viewer*

A web-based viewer has been developed to display the results of the algorithm using Django (Django, 2015) and Protein Viewer (PV) (Biasini, 2015). An interactive depiction of each of the grids has also been developed to allow the user to place a conformer inside the grid and investigate how the conformer overlaps with the grid. This allows users to visually inspect which parts of the conformer do not overlap with the grid and are being counted as novel. A visualisation has also been developed that shows the individual score of each atom or entity to allow the user to determine which atom, pharmacophoric feature or interactions are novel.

#### *2.2.7.1 Results Page*

The web-based results page is shown in Figures 2.10 and 2.11. Figure 2.10 shows the left-hand side of the screen, and Figure 2.11 continues from this as if the user had scrolled to the right. A depiction of each candidate conformer is shown down the left hand-side (Figure 2.10). For each score, the results page shows the average score for all the conformers and a histogram depicting the number of conformers with a given score. The histograms should allow the user to make well-informed decisions, as the

average score does not capture the spread of scores achieved by the number of conformers. If all the conformers achieve a very similar score, then the average may be an adequate representation. However, if some conformers are extremely novel and some are much less so the highest or lowest score may be more useful than the average. The scores are shown in order of Element Score, Pharmacophore Score and then Interaction Score. The table can be toggled to sort the compounds by the average Element, Pharmacophore or Interaction Score, and can be sorted in either ascending or descending order.

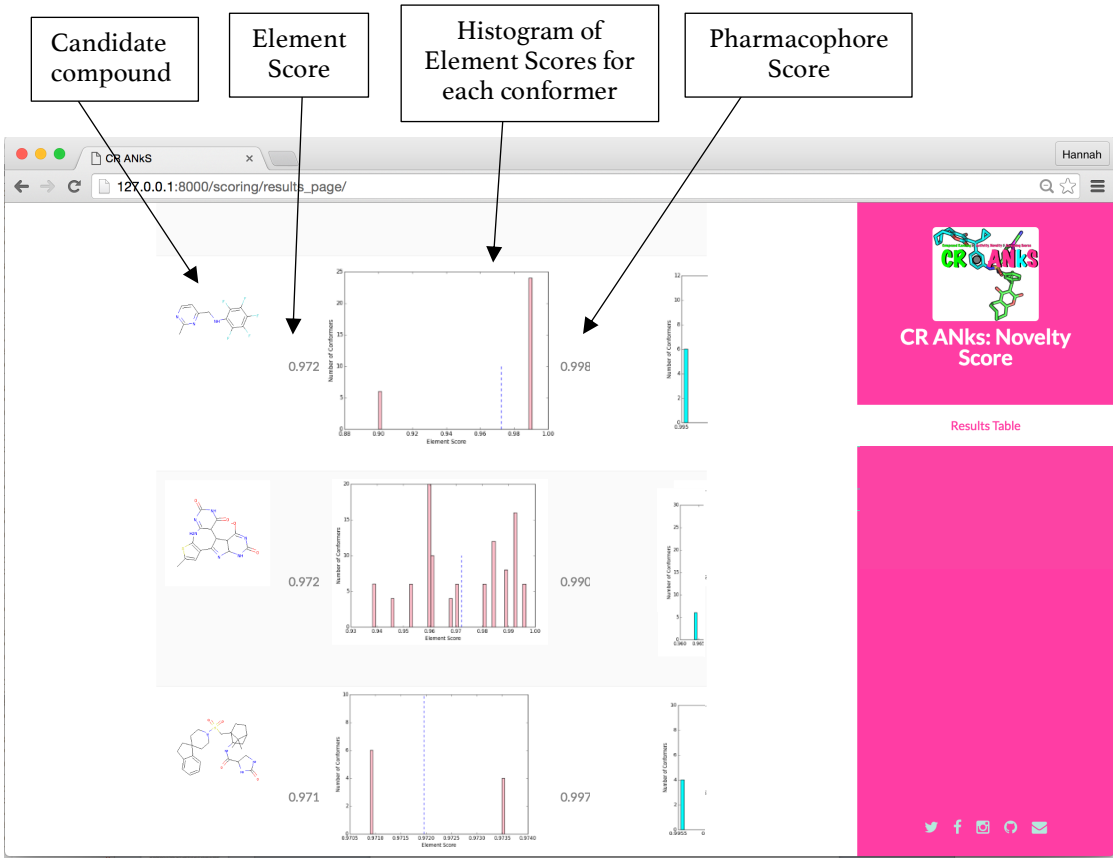



Figure 2.10 Results Page of the CRANkS Algorithm, part 1. A depiction of the candidate compound is shown on the left, with the Element Score, a histogram of the Element Scores for each of the conformers and the Pharmacophore Score. These are labelled. The results page continues in Figure 2.11 as if the user has scrolled to the right. The compounds are the HIV-1 protease ligands used as candidate compounds from data set 1 discussed in Section 2.3.2.

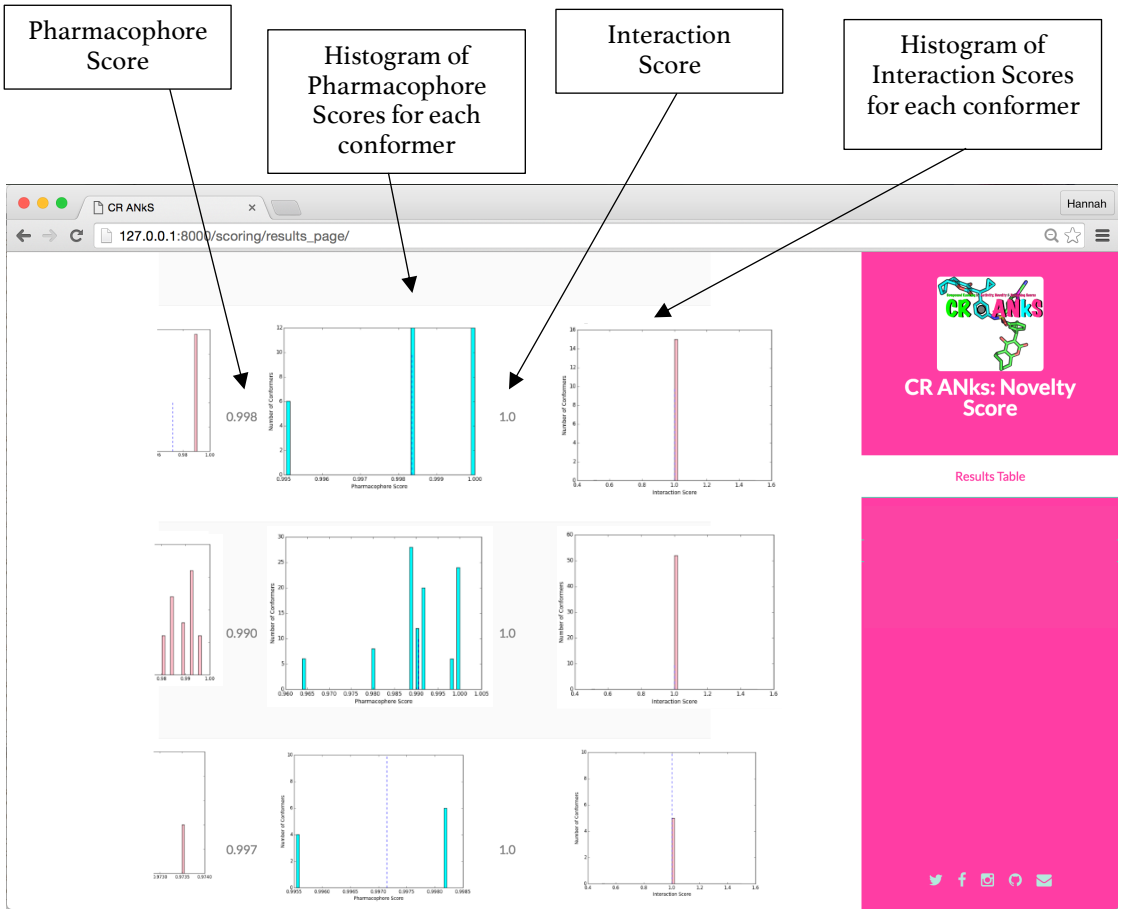



Figure 2.11 Results Page of the CRANkS Algorithm, part 2. This figure continues from Figure 2.10 as if the user has scrolled to the right. The Pharmacophore Score, Interaction Score, and histograms of each of these scores for each of the conformers of a given candidate compound are shown. These are labelled. The compounds are the HIV-1 protease ligands used as candidate compounds from data set 1 discussed in Section 2.3.2.

### *2.2.7.2 The Element Grid*

The web-based viewer for the element grid is shown in Figure 2.12. The element grid is shown within the protein structures used for the calculations. The grid points are represented by transparent spheres. The size of the sphere indicates the count at that grid point and the colour of the sphere indicates the element type being counted. The grid points can be shown or hidden to allow the grid to be further investigated.

For example, the carbon grid points can completely dominate the grid, so by hiding those points the other types of elements can be seen more clearly.

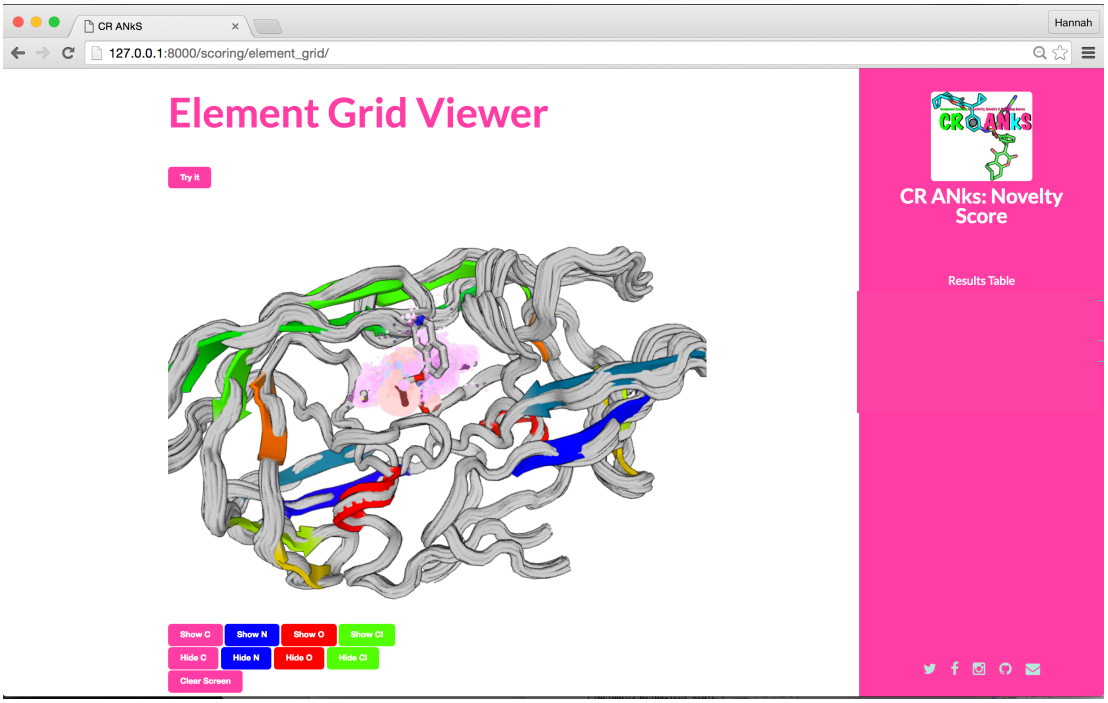



Figure 2.12 Web-based viewer to show the CRANkS element grid. The grid is displayed using transparent spheres: the size of the sphere indicates the element count at that point, the colour indicates the element. Spheres of a certain element can be toggled on/off using the buttons at the bottom of the viewer. A conformer is shown inside the grid - any conformer can be loaded into the grid using the conformer id number, refered to as the primary key. Carbon atoms are represented by pink spheres, nitrogen by blue, oxygen by red and chlorine by green.

Each conformer of the candidate compound can be loaded into the grid. This allows the user to determine which parts of the conformer are occupying the same points as the prior knowledge but also clearly shows which parts of the molecule are being counted as novel. This is also aided by another visualisation tool as shown in Figure 2.13. The conformer can be shown with spheres representing the grid point each atom is assigned to. The colour of the sphere indicates the novelty score at that point. This indicates which atoms are being scored as novel and which atoms are not, to

help the user understand the Element Score for the conformer and make informed decisions.

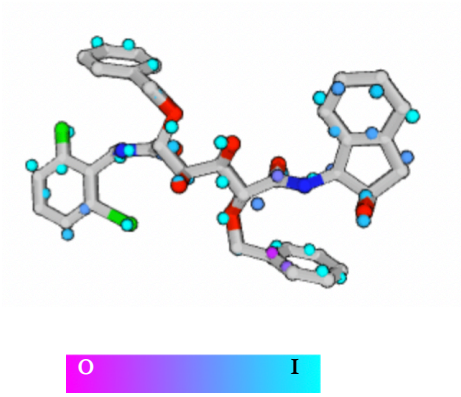



Figure 2.13. Element Score per atom for a compound. Novelty increases from 0 to 1, with colour change magenta to cyan respectively.

### *2.2.7.3 The Pharmacophore Grid*

The CRANkS pharmacophore grid is also represented by transparent spheres in a web-based viewer as shown in Figure 2.14. Each sphere represents a grid point - the size of the sphere indicates the count of pharmacophoric features at that point, and the colour indicates the type of pharmacophoric feature being counted. The different types of pharmacophoric features can be toggled on or off using the buttons below the grid to enable the user to see the grids of different features more clearly. Conformers can also be loaded into the grid using the conformer primary key to investigate how the conformer overlaps with the grid. This allows the user to understand the score assigned to the candidate compound, and make informed decisions about whether the candidate compound should be taken forward to be synthesised.

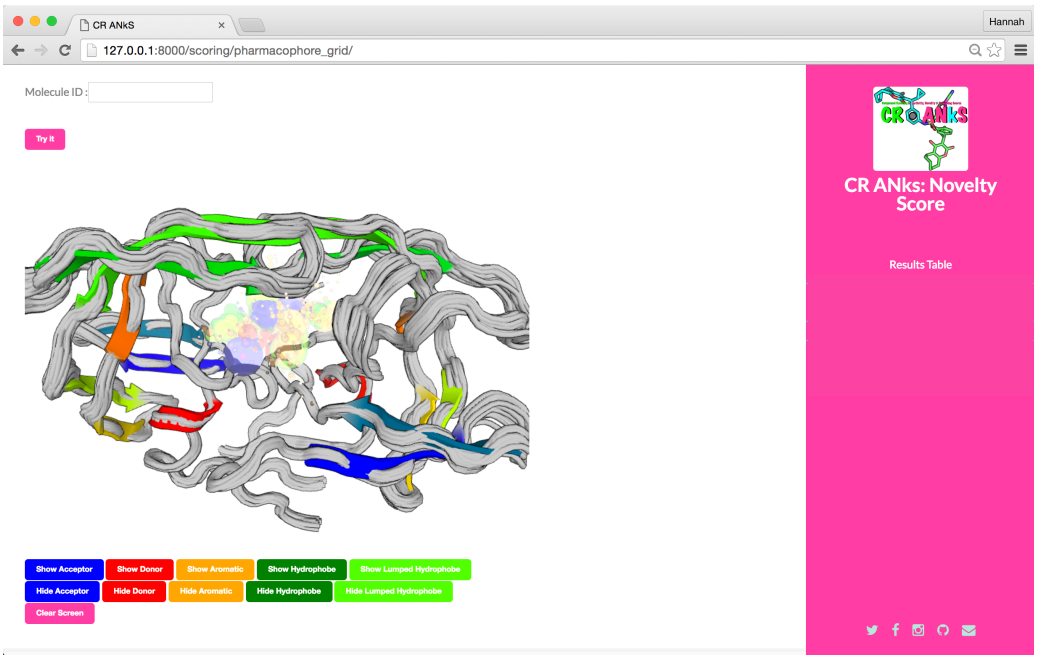



Figure 2.14. Web-based viewer to show the pharmacophore grid. The grid is displayed using transparent sphere: the size of the sphere indicates the pharmacophoric feature count at that point, the colour indicates the type of pharmacophoric feature. Acceptor features are shown in blue, donor in red, aromatic in orange, hydrophobe in green and lumped hydrophobe in light green. Spheres of a certain type can be toggled on/off using the buttons at the bottoms of the viewer.

### *2.2.8.4 The Interaction Grid*

The web-based viewer depicting the CRANkS interaction grid is shown in Figure 2.15. Each interaction counter that forms the interaction grid is shown as a cylinder. The centre of each face of the cylinder is either the start or end grid point. The radius of the cylinder is proportional to the count of the interactions between the two grid points and the colour indicates the type of interaction. Interaction counters of a particular type can be hidden or shown using the buttons below the viewer. Hydrophobic interactions usually dominate the grid, so it can be useful to hide these if other parts of the grid are being examined. Conformers can be loaded into the grid, using the primary key of the conformer (the id number used in the CRANkS

database). Interaction counters that occupy the same starting point as the interaction are shown as opaque. This allows the user to see which interactions are conserved between the prior knowledge and the conformer, but additionally which interactions are not satisfied by the conformer but appear in the grid and also interactions that the conformer forms that are novel compared to the grid.

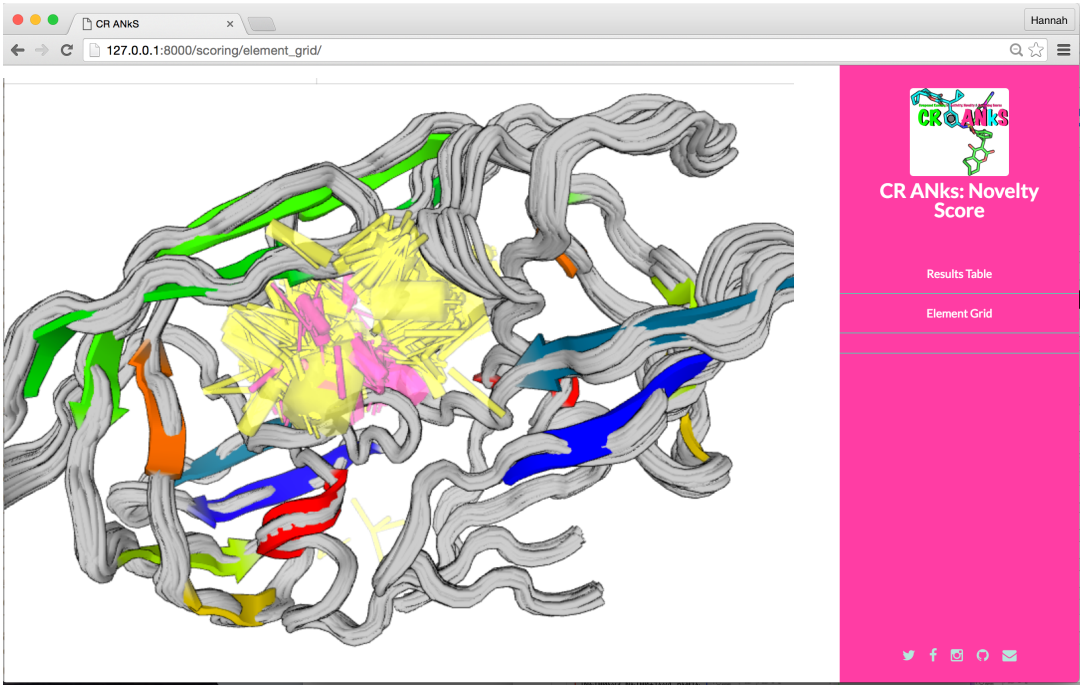



Figure 2.15. Web-based viewer showing the interaction grid. Interaction counters that form the grid are displayed using transparent cylinders: the radius of the cylinder indicates the interaction count at that point. The colour indicates the interaction type. Yellow cylinders show hydrophobic interactions, pink show hydrogen bonds.

### *2.2.8 Database Schema*

All the data, each step of the algorithm and the results are stored in a MySQL database. This allows the results to be easily accessible using Django (Django, 2015) to display the results as a web page. It also allows each step to be calculated separately as the database can be accessed for information sequentially. The results

of each step can also be accessed once the algorithm is completed to investigate the method. The structure of the database or database schema is shown in Figure 2.16. The database is split into two sections. The fields that store information for the structural data and generated grids are described in the *data_input* data model and are shown in the bottom half of the figure. Fields that store information for the candidate compounds and scores are described in the *scoring* data model, in the top half of the figure.

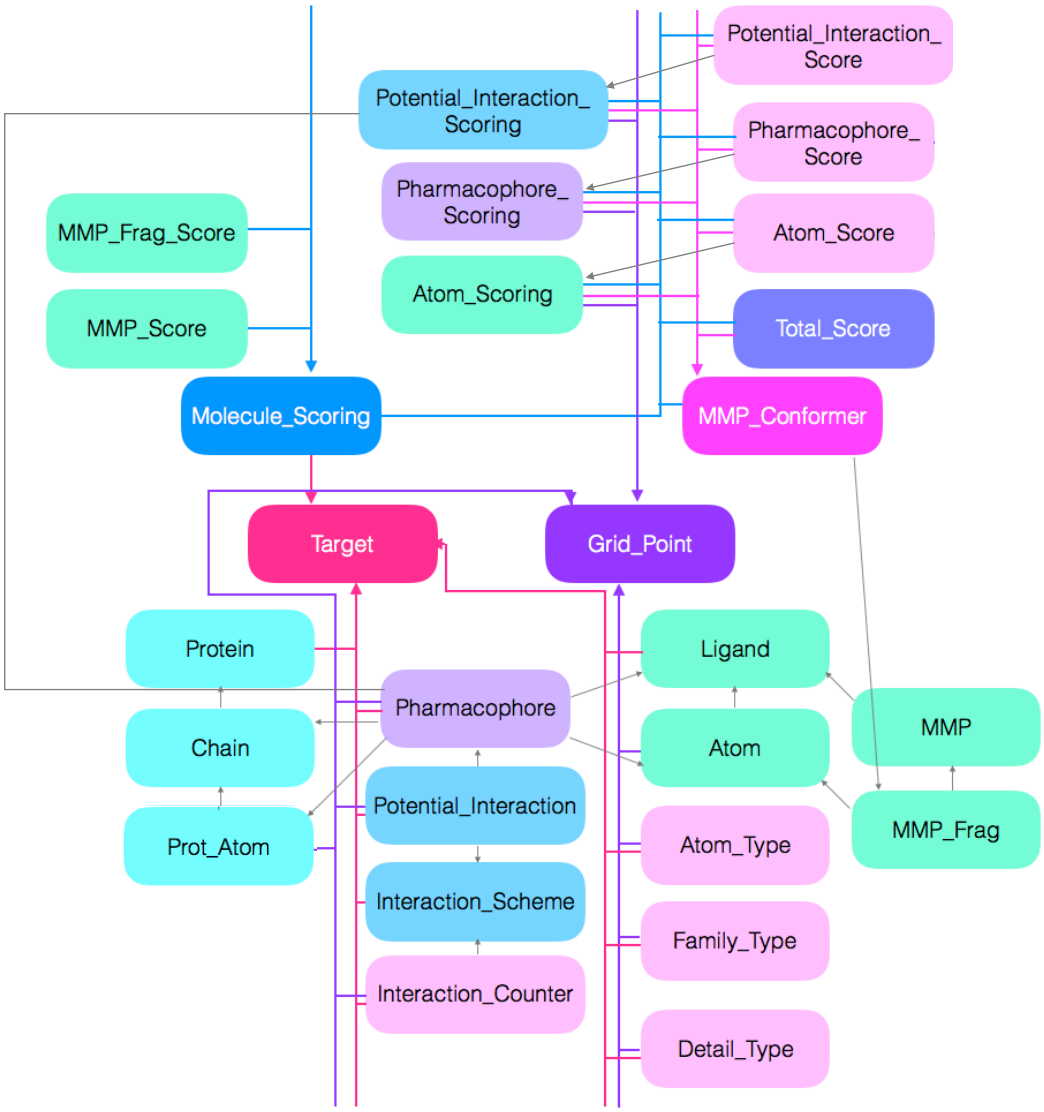



Figure 2.16. Database schema used to store each step and the results of the algorithm. Models are shown in rectangles. Foreign keys are depicted using arrows.

### *2.2.8.1 Data_input Models*

The central object in the *data_input* model is the *Target*. All other objects are linked either directly or indirectly through multiple foreign keys to the *Target* object. It allows all the data for a particular project to be found. The *Target* is defined by a target name which must be unique to the project. The *Protein* and *Ligand* objects are linked to the target by a foreign key. The other central object in both the *data_input* and *scoring* models is the *Grid_Point*. *Grid_Point* objects store the x, y and z coordinates of each individual grid point - one object per grid point. The coordinates of any part of the protein-ligand system are then described through foreign keys to the corresponding *Grid_Point* object.

The *Ligand* object stores the SMILES and 3D coordinates of each ligand from the protein-ligand complexes after the ligand is treated as described in Section 2.3.2. Information about each atom of each ligand is then stored in the *Atom* object. The nearest grid point to each atom is linked to the atom object by a foreign key to the corresponding *Grid_Point* object. The parent ligand is also linked by a foreign key. Each atom is stored both to generate the element grid, but also to indicate which atoms are involved in a particular pharmacophoric feature. Each ligand is also broken into building blocks as part of the 3D Matched Molecular Pairs process described in Section 2.6.1. For each bond broken, the two resultant blocks are stored in the *MMP* object, which is linked to the parent ligand by a foreign key. Each unique building block is then also stored in the *MMP_Frag* object, with the coordinates for the block.

The *Protein* object stores a copy of the PDB file defining the coordinates of each protein. Each protein chain is then stored separately as a *Chain* object after the processing described in Section 2.3.1. The *Chain* object is linked to the *Protein* object by a foreign key. Each atom in the protein is then stored in the *Prot_Atom* object. The nearest grid point to the protein atom is linked by a foreign key as well as the parent chain. Each protein atom is stored to indicate which atoms make up a pharmacophoric feature and also to allow conformers that occupy the same grid point as a protein atom, i.e. clash with the protein, to be identified.

Pharmacophoric features for both the protein and the ligand are stored using the *Pharmacophore* object. This holds the type of the feature, whether the feature is found on a ligand or pharmacophore and is linked to the corresponding *Ligand* or *Protein* object. The object is linked to the atoms that make up the feature by a many-to-many key to the *Prot_Atom* and *Atom* objects. It also is linked to the grid point corresponding to the location of the feature: in the case of aromatic rings, for example, this is the centre of the ring.

Protein-ligand interactions are stored as *Potential_Interaction* objects. The object stores the type of interaction and points to the P*harmacophore* objects that make up the interaction. It is also linked to the grid points that make up the interaction via a many-to-many key to the corresponding *Grid_Point* objects. Interaction schemes that define interactions, in terms of interaction type, which features make them, distance and angles are stored using the *Interaction_Scheme* object. Interactions are then calculated using an *Interaction_Scheme* object, and the *Potential_Interactions* stored are linked to the scheme via a foreign key.

The *Atom_Type* object is used to store the points of the element grid. The object stores the type of element and the count of that element for the given grid point. It then points to the *Grid_Point* using a foreign key. The *Family_Type* object is used to store the points of the pharmacophore grid. The type of pharmacophoric feature and the count of the feature at that grid point are stored, as well as a foreign key to the corresponding *Grid_Point* object. The *Detail_Type* stores an alternative version of the pharmacophore grid using the more detailed feature classification of RDKit as discussed in Section 2.4.2. The *Interaction_Counter* object stores the interaction counters that make up the interaction grid. The type of interaction and the count are stored, as well as the primary keys of the starting and end grid points. The corresponding grid points are also linked to the object by a many-to-many key.

#### *2.2.8.2 Scoring Models*

The scoring models are used to store data related to the candidate compounds. The *Molecule_Scoring* model is central to this system. This stores initially just the SMILES of the candidate compound, and the path to an image of the compound (for use in the web-page viewer). However, eventually a *Molecule_Scoring* object also holds the total Element Score, Pharmacophore Score and Interaction Score for the candidate compound, as well as the ratio of heavy atoms for the conformer generated part of the molecule, to the heavy atoms of the constrained part. Additionally, paths to histograms displaying the three scores for each generated conformer, and to an SDF file containing the conformers are stored (all of which are used in the web-based viewer). *Molecule_Scoring* objects are linked to the corresponding *Target* object with a foreign key. The *MMP_Conformer* objects are also central to the data models and hold

information for the conformers generated for each candidate compound. It is linked to the corresponding *Molecule_Scoring* object via a foreign key. *MMP_Conformer* objects store the conformer as a *mol_block* and also point to the corresponding *MMP_Frag* from the *data_input* models, that was used in the constrained conformer generation.

The candidate compound is split into building blocks to enable Matched Molecular Pairs to be found, as discussed in Section 2.6.1. The resultant blocks from each bond break are stored together as an *MMP_Score* object, with a foreign key to the parent *Molecule_Scoring* object. Each individual block is then stored as an *MMP_Frag_Score* object which points to the corresponding *MMP_Score* object.

Each atom of each conformer is stored as an *Atom_Scoring* object. This holds the type and points to the nearest grid point via a foreign key to the corresponding *Grid_Point* object. The pharmacophoric features of each conformer are stored as a *Pharmacophore_Scoring* object, which holds the type of feature and is linked to the atoms that make up the feature by a many-to- many key to the corresponding *Atom_Scoring* objects. The *Pharmacophore_Scoring* object also points to the grid point corresponding to the centre of the feature by a foreign key to the corresponding *Grid_Point* object. The *Potential_Interaction_Scoring* object stores each individual protein-ligand interaction formed by the conformer. The type of interaction is stored, and the object is linked to the pharmacophoric features that make up the interaction using a foreign key to the corresponding *Pharmacophore_Scoring* object of the conformer, and the *Pharmacophore* object of the protein. The *Potential_Interaction_Scoring* object also stores the primary key of the *Grid_Point* object

referring to the starting grid point, and the primary key of the end *Grid_Point* object. A foreign key also links it to the *Interaction_Scheme* object used to define the interaction. The *Atom_Scoring*, *Pharmacophore_Scoring* and *Potential_Interaction_Scoring* objects are all linked to the parent conformer and candidate compound by foreign keys to the *MMP_Conformer* and *Molecule_Scoring* objects.

The *Atom_Score* object holds the Element Score for a given individual atom and is linked to the *Atom_Scoring* object corresponding to the atom being scored by a foreign key. The *Pharmacophore_Score* object stores the Pharmacophore Score for an individual pharmacophoric feature and is linked to that feature via a foreign key to the corresponding *Pharmacophore_Scoring* object. The *Interaction_Score* object holds the Interaction Score for an individual protein-ligand interaction formed by a conformer and the corresponding *Potential_Interaction_Scoring* object is linked by a foreign key. Total Score objects hold the total element, pharmacophore or Interaction Score for a given conformer. The type of score is indicated by a key field that reads either 'element', 'pharmacophore', or 'interaction'. The *Atom_Score*, *Pharmacophore_Score*, *Interaction_Score* and *Total_Score* objects are linked to the parent *MMP_Conformer* and *Molecule_Scoring* objects by foreign keys.

## 2.3 Preliminary Results

### *2.3.1 HIV-1 Protease*

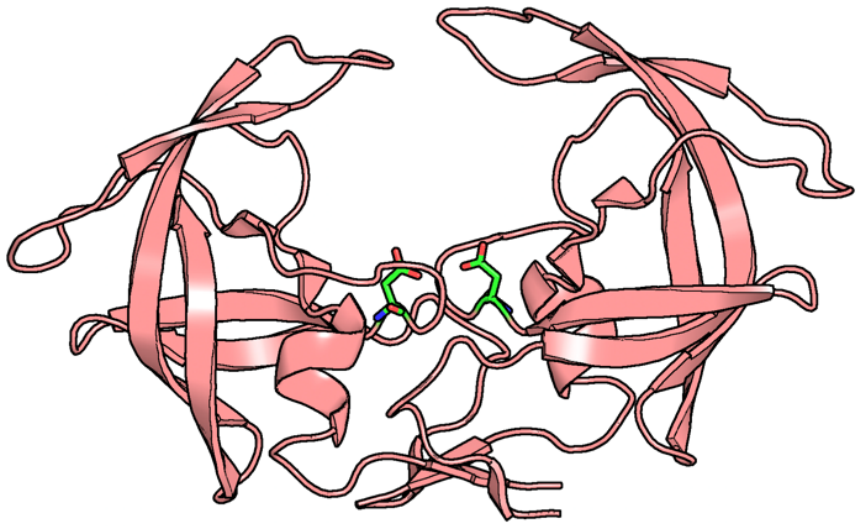


Figure 2.17 The apo-form of HIV-1 protease (PDB structure 2r8n; Coman *et al.,* 2008). The two catalytic aspartate residues at position 25 are shown in green. Produced using PyMOL ( Schrödinger, LLC.)

HIV-1 protease was chosen as the target for preliminary testing of the CRANkS algorithm. The homodimeric enzyme is essential for the life cycle of the HIV virus: the protein cleaves polyproteins that are required for the virus to replicate. Consequently, there has been much research to identify HIV-1 protease inhibitors, and this has led to large amounts of experimental data available for binders to the target (Brik and Wong, 2008).

HIV-1 protease exists as a homodimer of two identical chains shown in Figure 2.17. The binding site is defined by residue Asp-25 on each chain and forms the catalytic

active site. Additionally, there is a flap on each chain which effectively opens or closes moving up to 7 Å upon binding with the substrate.

Although there is a lot of publicly available experimental data, one potential problem with using HIV-1 protease is the high mutation rate of the protein and the emergence of drug-resistant mutants. Therefore, care must be taken to ensure the protein-ligand complexes have identical sequences, as binding affinity can be dramatically affected by a single amino acid mutation.

## *2.3.2 Pre-processing of Data Set 1*

### *2.3.2.1 Structural Data*

To generate the CRANkS grids, structural data is required. All the HIV-1 protease protein-ligand complexes were downloaded from PDBbind-CN (Liu *et al.*, 2015) - a database of protein-ligand complexes found in the Protein Data Bank (Berman *et al.*, 2000) with associated binding affinities. The complexes were clustered by 100% sequence identity and the cluster matching the sequence of the largest set of actives from BindingDB (Gilson *et al.,* 2016) was taken forward. This consisted of 59 protein-ligand complexes, which are listed in Appendix B, Table B.2.1.

The protein-ligand complexes were then aligned using the *align* function in PyMOL (Schrödinger, LLC.), as shown in Figure 2.18. The complexes were aligned using the backbones of residues 20 to 30 on both chains (shown in Figure 2.18). This is centered around the catalytic residue at position 25 so should ensure the binding

sites of the complexes are aligned. The complexes appear well-aligned with little difference in the protein conformations (Figure 2.18). The average RMSD for residues 20 to 30 using the backbone was 0.27 Å and the maximum RMSD between any two structures for this selection was 0.69 Å. The average RMSD for residues 20 to 30 for all atoms was 0.65 Å, and the maximum RMSD was 1.16 Å.

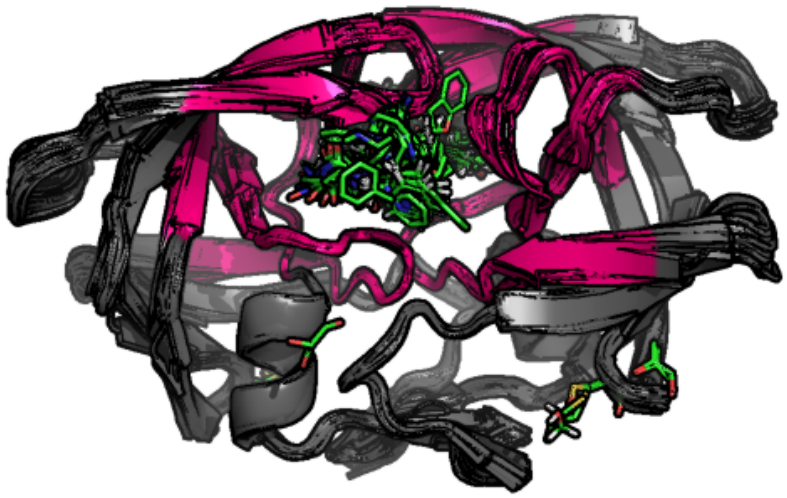


Figure 2.18 Protein-ligand complexes of HIV-1 protease used to construct CRANkS grids for data set 1. The binding site, defined as residues 20 to 30, is coloured in pink. The structures were aligned using the backbone atoms of this region on both chains. The structures are listed in Appendix B, Table B.2.1.

### *2.3.2.2 Candidate Compounds*

To investigate the CRANkS algorithm a set of candidate compounds are required to be scored. I decided to test the algorithm on a set of actives and inactives, to investigate how both groups are scored and because in hit-to-lead optimisation the algorithm will be testing a set containing a mixture of active and inactive compounds. Although CRANkS was devised for novelty the novelty of the compound could be linked to its activity by the similarity property principle – if the

compounds are more similar to the ligands the compounds could be more likely to be active. Thus the ability of CRANkS to separate actives and decoys was tested.

The active compounds were curated from the BindingDB database (Gilson *et al.*, 2016). All compounds from the database with activity against HIV-1 protease were downloaded along with the sequence of the target the compound was tested against. The compounds were then grouped by 100% sequence identity, and the group of compounds bound to the HIV-1 protease that matched the sequence of the structural data (in Section 2.3.2.1) were taken forward as the active compounds. This consisted of 226 molecules. First molecules are desalted. A chirality filter was then applied to remove molecules with ambiguous chiral centres. This is to avoid issues with stereoisomers if there is not information as to which stereoisomer has been tested, or whether it was a racemic mixture. 116 molecules remained after this filter. Next a PAINS filter was applied to try and remove any false positives. No molecules were filtered out by the PAINS filter. The 116 molecules were then used as the active molecules in the test set of candidate compounds. The SMILES of these molecules are listed in Appendix B, Table B.2.2.

The remaining set of 116 active molecules were then investigated in terms of how much of the molecule is found to match with its Matched Molecular Pair within the set of ligands used to create the grid. To place the matched candidate compound within the binding site a constrained conformer generation is performed as described in Section 2.2.5. Part of the molecule that matches a part of a ligand from the structural data is constrained. One hundred conformers are generated for the rest of the molecule. To investigate whether this would also have an effect on the

result, the ratio of the number of heavy atoms of the flexible part to the number of heavy atoms in the whole compound is plotted as a histogram for the set of molecules as shown in Figure 2.19. There are two clear groups above and below approximately 0.72.

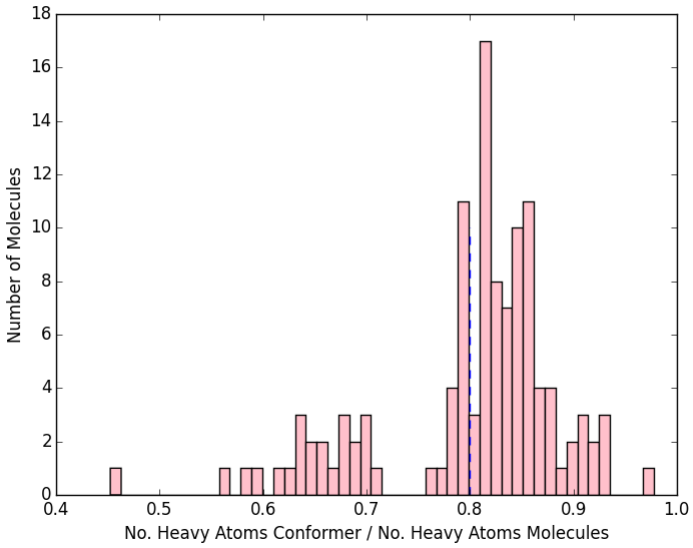



Figure 2.19 Histogram of ratio of the number of heavy atoms in the part of the molecule that a conformer is generated for, to the number of heavy atoms in the molecule for a set of active binders for HIV-1 protease. The blue dotted line is the mean.

A set of inactive compounds, or decoys, was required as a comparison to the set of active compounds. This set was curated from the DUD-E database (Mysinger *et al.,* 2012) which compiles benchmark test sets of binders and putative non-binders. The set of decoys for HIV-1 protease were downloaded from DUD-E containing over 36,000 molecules. This was too large a set to be used for preliminary results testing. Therefore, a selection of the 500 most diverse compounds were taken to be used as a set of inactive compounds. These were selected using the MaxMin Algorithm and using Morgan Fingerprints as the descriptors as implemented in RDKit (Landrum,

2015, Ashton *et al.*, 2002; see Section 3.2.5.1). These decoys were subjected to the same pre-processing leaving 469 decoys for the dataset. The SMILES of the decoy compounds are listed in Appendix B, Table B.2.3.

This set of 469 diverse decoy compounds was then investigated in terms of whether 3D Matched Molecular Pairs could be found with the active ligands from the structural data. When the compounds are placed in the binding site using constrained conformer generation, part of the query molecule that matches part of a bound ligand is constrained and given the coordinates of that part of the ligand - the constrained part. For the rest of the compound, 100 conformers are generated - the flexible part.

A histogram of the number of heavy atoms in the flexible part as a fraction of the total number of heavy atoms of the molecule for the decoys is shown in Figure 2.20. When compared to Figure 2.19 there is a difference between the actives and inactives in terms of the distribution of the ratio, which may have an effect on the results. The size of the constrained part of the molecule could affect the accuracy of the conformer formed. If the active molecules inherently have a larger matched molecular pair than the inactives this could cause a bias towards a better separation of actives from inactives, due to the larger match the actives would be calculated to be more similar to the CRANkS grids. This artificial bias is unlikely to be the case in a real drug discovery campaign as hit-to-lead compounds would be much more similar.

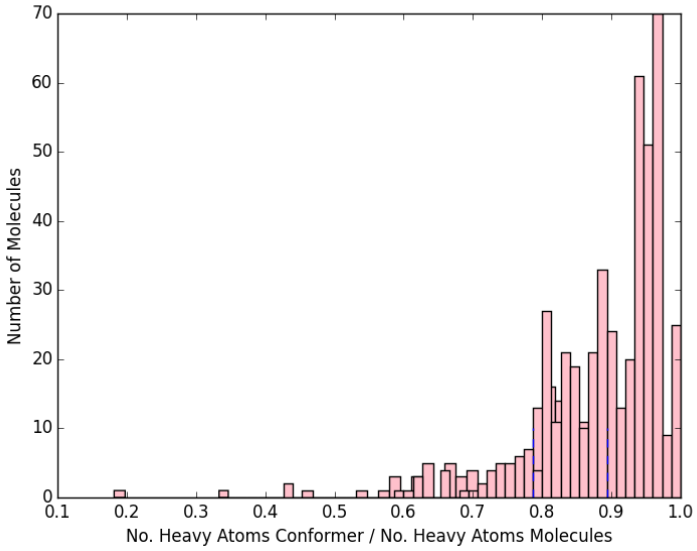



Figure 2.20 Histogram of ratio of the number of heavy atoms in the part of the molecule that the conformer would be generated for, to the number of heavy atoms in the molecule for a set of putative inactives for HIV-1 protease.

### *2.3.3 Discrimination between Actives and Inactives Set 1*

By scoring both active and inactive molecules I can investigate whether the CRANkS scores differentiate between the actives and inactives. The score has been developed for novelty and so should not necessarily discriminate - an inactive molecule can be just as novel as an active one and vice-versa. However, it is interesting to investigate whether the grids offer any additional information as to whether a molecule is likely to bind.

ROC (receiver operating characteristic) curves (see Section 3.2.3) have been plotted to show whether the Element and Pharmacophore Score is better than random at discriminating between binders and non-binders as shown in Figures 2.21. and 2.22.

These are compared to using a Tanimoto dissimilarity score – the Tanimoto coefficient (Section 1.3.1) between the candidate compound and each ligand in the set used to build the grid is averaged over the ligands, using Morgan Fingerprints (Section 3.2.5.1).

Both the Pharmacophore Score and Element Score show enrichment - molecules with lower Pharmacophore and Element Scores are more likely to be active. However, the simple Tanimoto dissimilarity score is much better at discriminating. This is likely to partly be because of how the decoy set was constructed: the inactive molecule set is much more chemically diverse than the active set and consequently much less chemically similar to the active set. Further investigation is needed to determine whether the discrimination of the Pharmacophore and Element Score is due to this, although it is unlikely as the scores were not found to be correlated with Tanimoto similarity in Section 2.3.7. Testing on inactive compounds that are as similar to the grid ligands as the active compounds would help to investigate this. Indeed, chemotype bias is a known criticism of many benchmark sets used for evaluating virtual screening methods, including DUD-E.

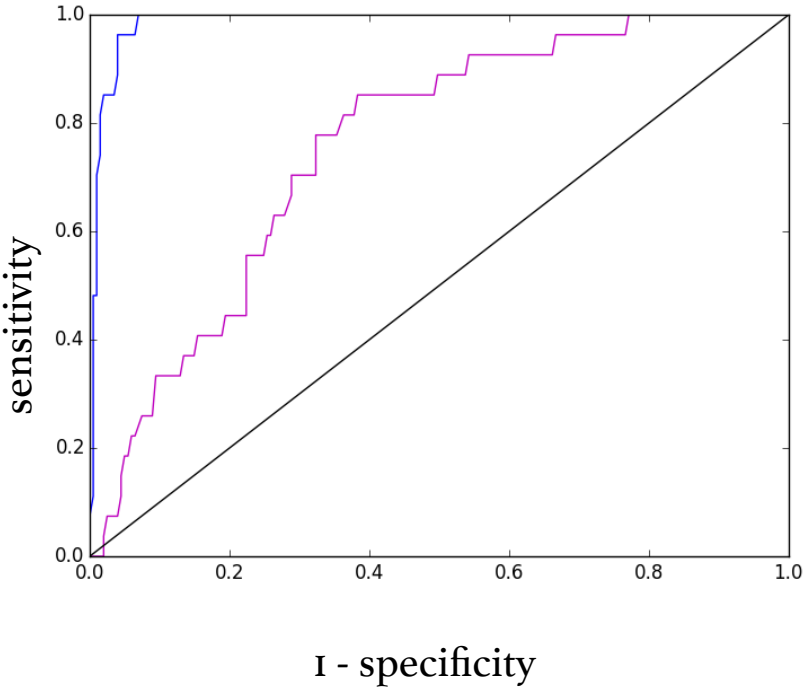



Figure 2.21. ROC curve of the Element Score, in pink, and Tanimoto score, in blue, for discriminating between actives and inactives for data set 1.

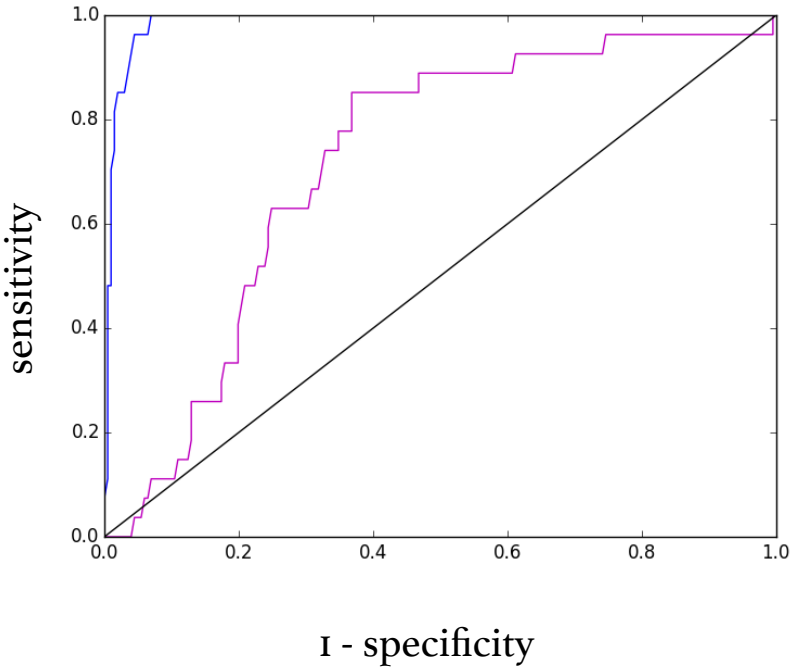



Figure 2.22. ROC curve of the Pharmacophore Score, in pink, and Tanimoto score, in blue, for discriminating between actives and inactives for data set 1.

### *2.3.4 Pre-processing of Data Set 2*

The structural data used to construct CRANkS grids was a small set of protein-ligand complexes found to have 100% sequence identity downloaded from PDBbind-CN (Liu *et al.*, 2015). A smaller subset of nine protein- ligand complexes were used to allow fast grid generation and to investigate parts of the method. A list of the PDB structures used, with the residue code for the ligand and resolution in each case, are shown in Appendix B, Table B.2.4. The complexes were aligned in the same way as the data in Section 2.3.2. This test set was chosen as the small number of molecules could be representative of the likely amount of data available in early hit-to-lead optimisation, where there may not be many structures of protein-ligand structures involving the target available.

The candidate compounds are a set of 59 molecules that were found to bind to HIV-1 protease with a different sequence from PDBbind CN. This is the same set used to construct grids in Data Set 1 (Appendix B, Table B.2.1) The candidate compounds were chosen as there are structures for the molecules available which would allow the conformer generation to be validated. Additionally, as the compounds are active against the same target with a differing sequence, there should be some similarities between the ligands used to create the grid and the candidate compounds.

### *2.3.5 Investigation of Conformer Generation (Set 2)*

To test whether the constrained conformer generation approach used by the algorithm produces accurate conformations, conformers were generated for a set of test molecules, for which there are X-ray crystal structures for the protein-ligand complexes. A box plot showing the RMSD (root mean square deviation) of the resultant conformers for each candidate compound are shown in Figure 2.23. There is clearly a large range of RMSD values from approximately 4 Å to 18 Å. No conformers are generated that are close to being within 2 Å of the crystal structure - an RMSD threshold that many docking algorithms currently achieve. This calculation does not subject the conformers to any further alignment to the crystal structure but is the RMSD between the conformer in its placement in the binding site and the crystal structure.

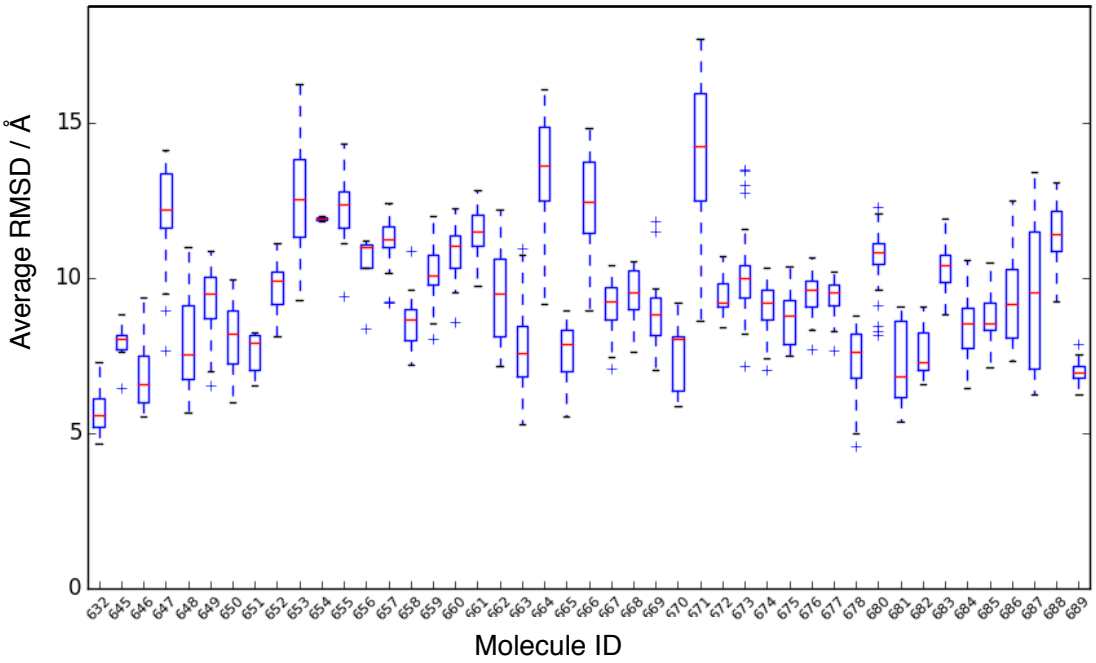



Figure 2.23 Box Plots to show the RMSD of the set of generated conformers for each candidate compound from the crystal structure.

To investigate how the accuracy of the conformer generation might depend on the properties of the small molecules, the correlation of the average RMSD versus a number of variables are plotted and shown in Figure 2.24. Figure 2.24 A shows the average RMSD plotted versus the number of conformers remaining once the conformers that sterically clash with protein atoms are deleted. Although initially 100 conformers are generated for all candidate compounds, the removal of clashing conformers causes a high variability in the number of conformers remaining. The figure shows little correlation between the final number of conformers and the average RMSD - the Pearson correlation coefficient for the values is -0.03. It is encouraging that despite the fact that the majority of conformers are deleted, this has little effect on the accuracy of the conformers to represent the actual conformation of the compound.

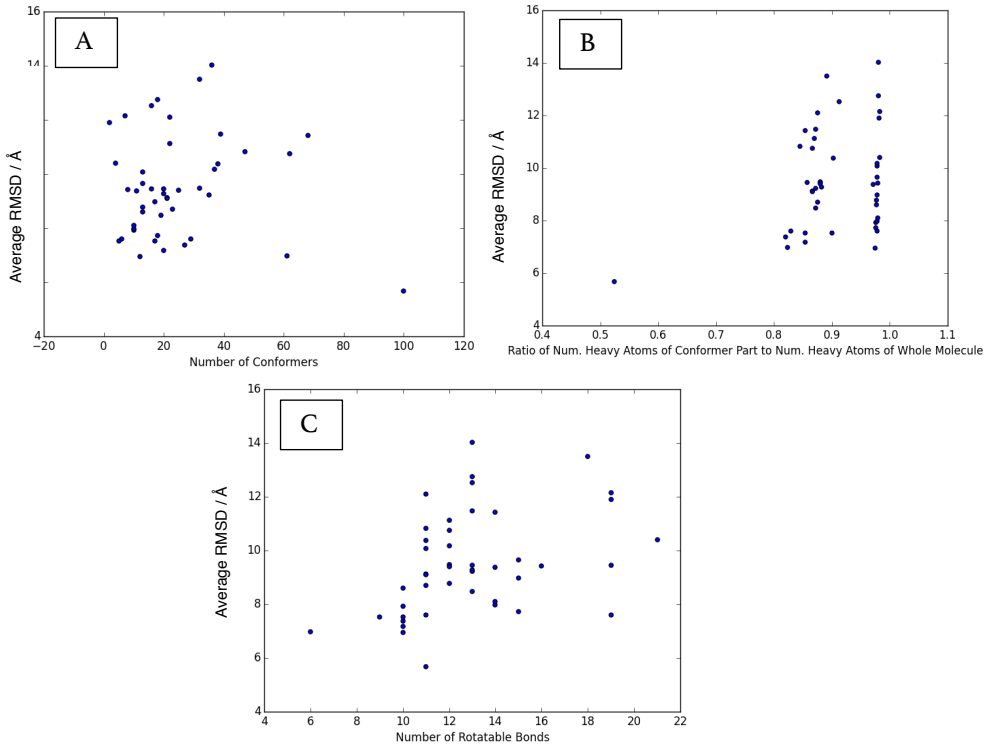



Figure 2.24 Average RMSD of the conformers generated for each candidate compound from the crystal structure plotted against (A) the number of conformers generated; (B) the ratio of the number of heavy atoms in the part of the molecule that had its conformation generated to the number of heavy atoms in the whole molecule; and (C) the number of rotatable bonds in the molecule.

Figure 2.24 B shows the average RMSD plotted against the ratio of the number of heavy atoms in the flexible part of the molecule versus the total number of heavy atoms in the molecule. As most of the conformers have a large ratio most of the molecules are flexible and do not benefit from retaining crystallographically observed binding modes in the small fraction held rigid.

The average RMSD of the conformers is also plotted against the number of rotatable bonds in the molecule in Figure 2.24 C, as this is known to affect the accuracy of conformer generation. There is some correlation and the Pearson coefficient was calculated to be 0.39.

Previous studies have found the RDKit conformer generation, in terms of reproducing the X-Ray crystal structure, to be correlated with the number of rotatable bonds, and was also able to generate conformers with an RMSD of less than 2 Å for molecules with even 12+ rotatable bonds (Ebejer *et al.,* 2012). However, this does not involve generation of the conformer within the binding site, and aligns the conformer to the crystal structure before calculating the RMSD. In our case, the conformer is not aligned to the crystal structure, as the calculation needs to reflect whether the two binding poses are close not only in the arrangement of atoms relative to each other but also placement in the binding site.

## *2.3.6 Scaffold-Hopping (Set 2)*

Scaffold-hopping is of high priority in hit-to-lead optimisation. By using experimental data, the knowledge-based CRANkS algorithm should be able to prioritise scaffolds that are different to the binders already seen but satisfy the same protein-ligand interactions. In this respect the ligand set of nine structures used to construct the CRANkS grids has five unique scaffolds as shown in Figure 2.25. The set of 59 candidate compounds has 33 unique scaffolds generated in Figure 2.26. The scaffolds were generated using RDKit by applying a Bemis-Murcko decomposition of the molecule (Bemis *et al.,* 1996). The scaffold was converted into a generic framework, removing all heteroatoms, and isomeric SMILES were used to match the scaffolds.

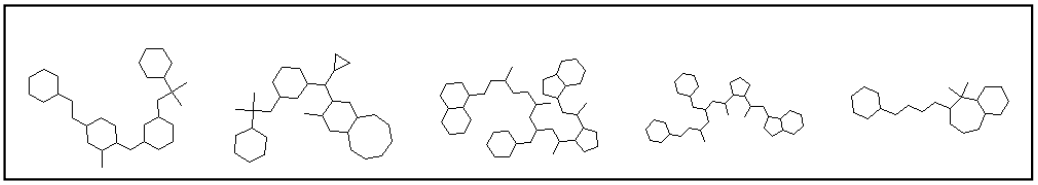


Figure 2.25 The five scaffolds calculated to describe the nine ligands used to create the CRANkS grids in data set 2.

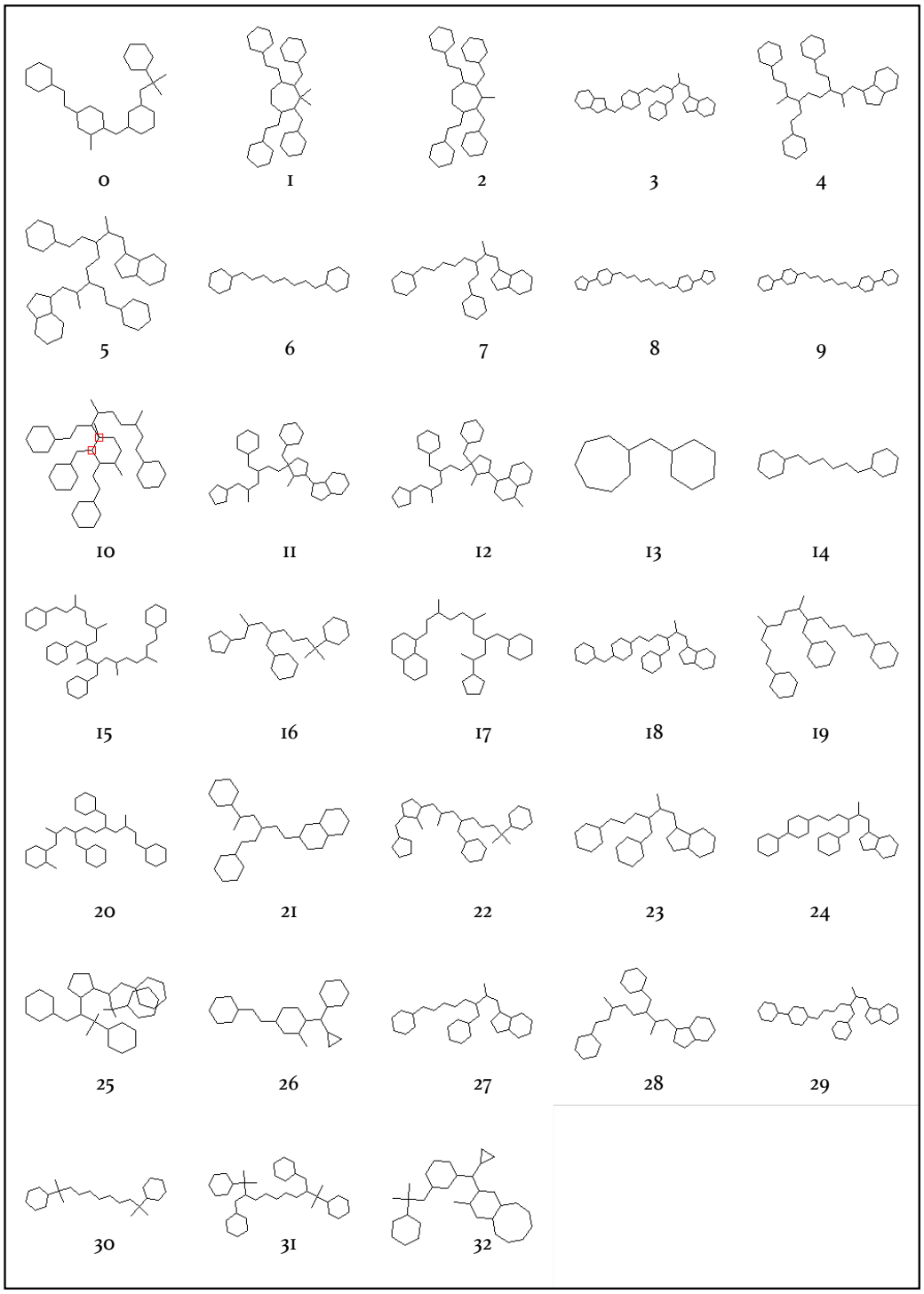



Figure 2.26 33 scaffolds calculated to describe the 59 candidate compounds scored by the algorithm in data set 2. The scaffold ID is shown below the scaffold.

There are some similarities between the scaffolds of the crystallographic ligands used to make the CRANkS grids and those of the candidate compounds. To investigate how each scaffold is ranked by the CRANkS algorithm, the novelty scores for each of the candidate compounds grouped by the calculated scaffold are plotted in Figure 2.27.

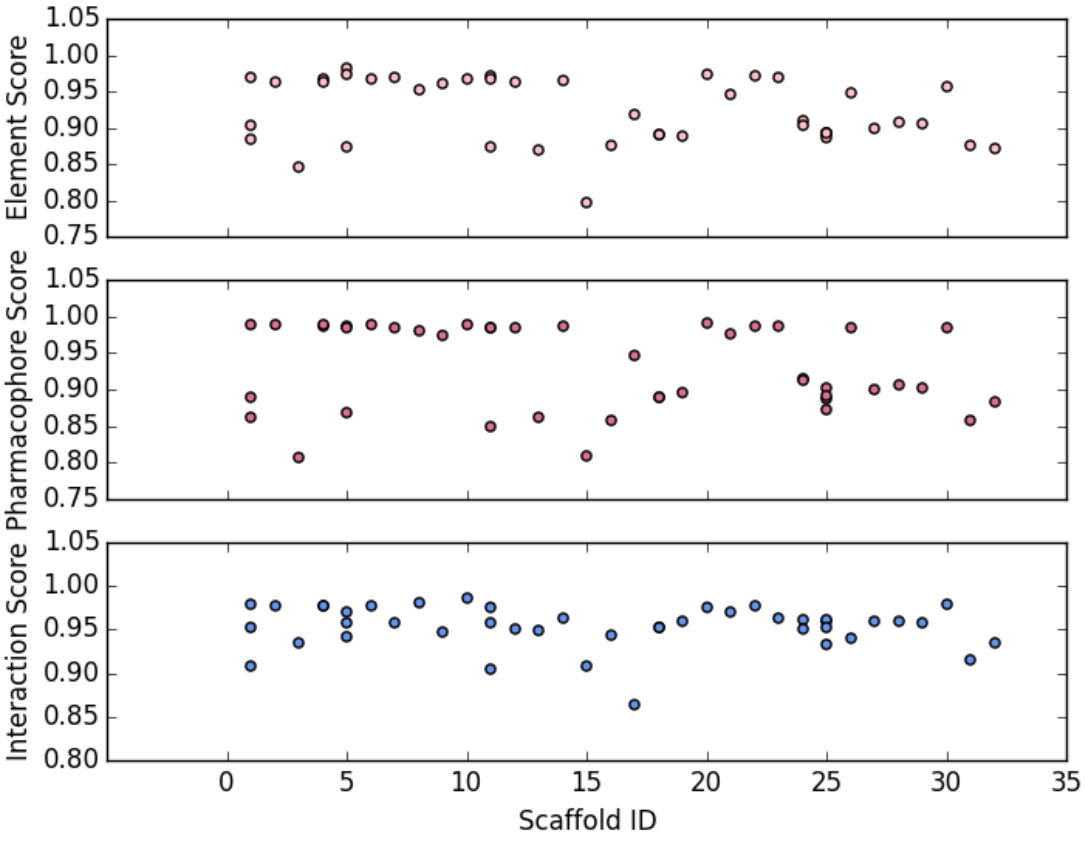



Figure 2.27. For each scaffold the Element, Pharmacophore and Interaction Score are plotted for each candidate compound that matches the scaffold.

Scaffold 32 matches a scaffold from the ligand set used to generate the grid. The corresponding Element, Pharmacophore and Interaction Score shown in Figure 2.27 indicate that the scores reflect this as the molecule has a low Interaction, Pharmacophore and Element Score. It is encouraging that the algorithm scores the scaffold low with all three scores - the molecule is found to satisfy interactions formed by known binders,  and pharmacophoric features and has significant overlap with the element grid.

Visualisations of the conformers of the candidate compound that is calculated to consist of scaffold 32 are shown in Figure 2.28. Figure 2.28 A shows the conformers with the interaction counters from the CRANkS grid shown as transparent cylinders, and the interactions of the conformers are opaque cylinders. There is a clear cluster of hydrophobic interactions (shown in yellow) which overlap with some of the hydrophobic interactions of the grid, which could account for the lower score. However, none of the hydrogen-bonds calculated match those from the grid. Figure 2.28 B shows the pharmacophoric grid using transparent spheres and the pharmacophoric features of each conformer as small opaque spheres. Notably, there is a large yellow hydrophobic sphere in the centre of the grid, which many hydrophobic features found on the conformers overlap with, which is likely to cause the low Pharmacophore Score. C and D shown the conformers within the element grid - C including the carbon spheres and D without the carbon spheres. Although an oxygen group can be seen to be in the same vicinity as a cluster of oxygen spheres from the element grid, there seems to be little other overlap other than a small part of the molecule overlapping with the carbon grid for most conformers.

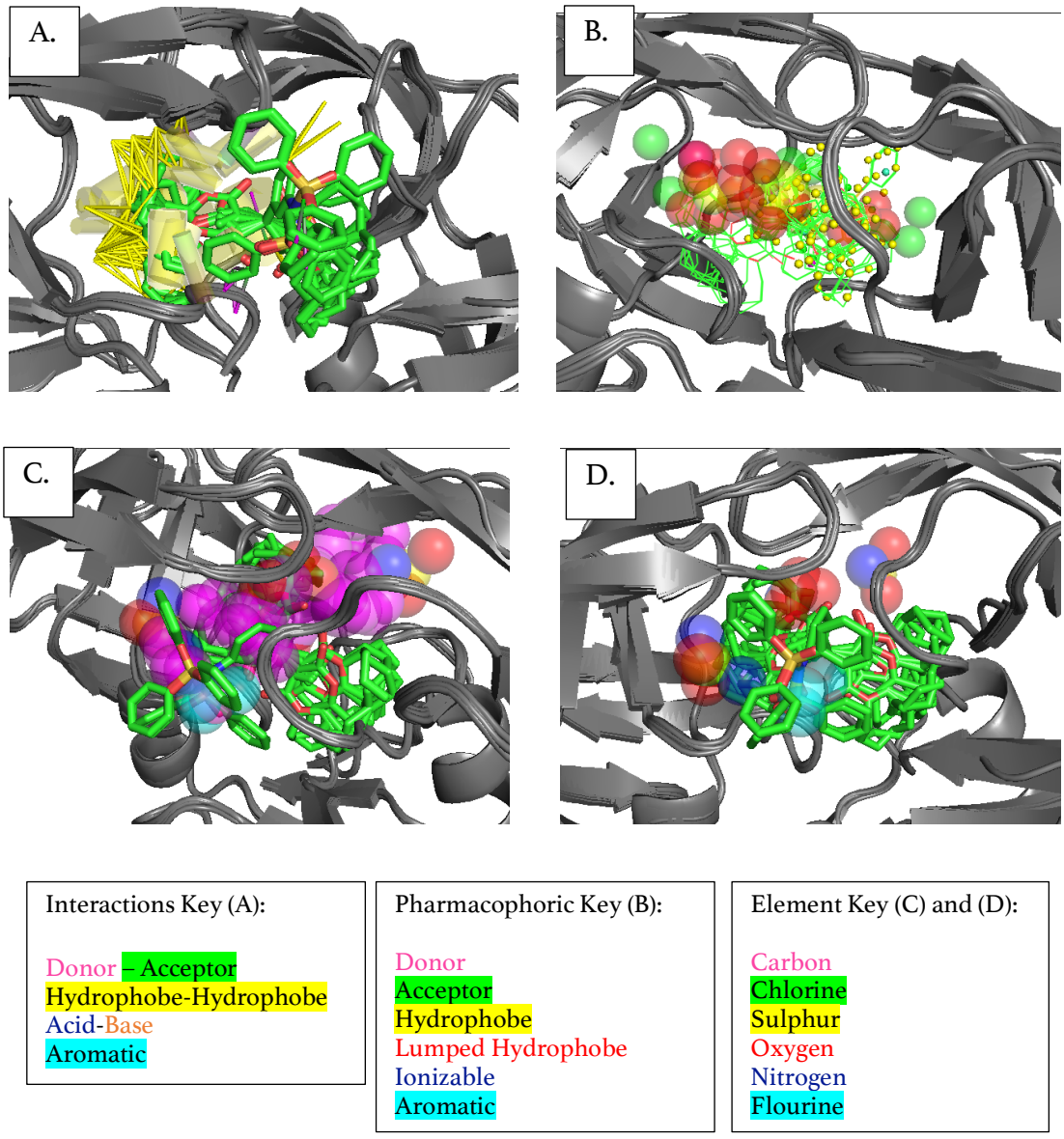



Figure 2.28 Conformers of a molecule that has the same scaffold as one of the ligands used to generate the grid shown in A: the interaction grid, with interaction counters shown transparent and interactions of the conformer opaque, B: the pharmacophore grid shown as transparent spheres with pharmacophoric features shown as opaque spheres, C: the element grid including carbon and D: the element grid without carbon. The size of the sphere or cylinder representing the grids indicates the count, and the colour indicates the type of feature. The molecule exhibits low Interaction, Element and Pharmacophore Scores which can be seen by the overlap of the conformers with the grids. Images created using PyMOL (Schrödinger, LLC).

All the CRANkS grids in Figure 2.28 show that the conformers are not within the same plane as the grid. The ligands used in the grid generation all reside along roughly the same axis, whereas the conformers generated span out into different directions. This is particularly noticeable in Figure 2.29. This shows the hydrophobic

parts of the pharmacophoric grid, with the ligands used to generate the grid. All the ligands reside in a similar volume. The grid shows two hydrophobic regions in line separated by a non-hydrophobic region. The non-hydrophobic region lies in line with the catalytic residues of the target. This pattern captured by the grid is likely to be important and should be mirrored by any candidate compounds prioritised for synthesis. Conformers that adopt this pattern should have a low Pharmacophore Score. However, if the conformers do not align with the orientation of the ligands and the grid then it is not possible to determine whether the molecule follows this pattern.

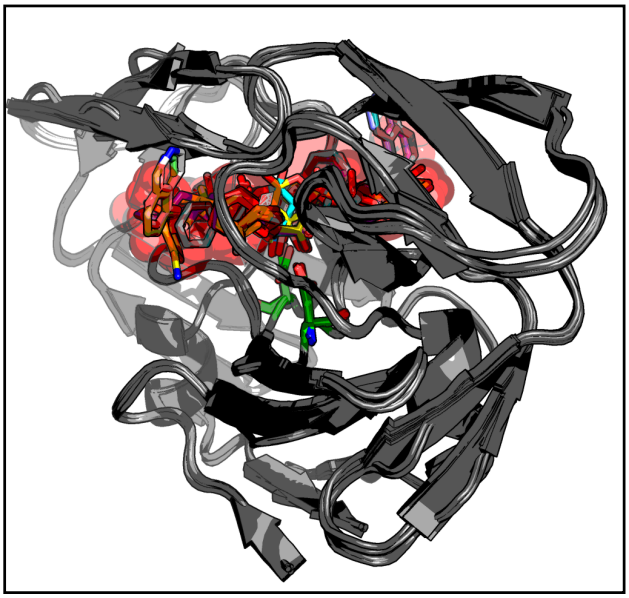


Figure 2.29. The hydrophobic part of the pharmacophore grid is shown by the lumped hydrophobe red spheres, with the ligands. There are two distinct areas with a split in the middle that is aligned with the catalytic residues.

Thus, it would be useful if the conformer generation could be further constrained so that the conformations fall within the same approximate area as the ligands used to create the grid. Generating more conformers would not be efficient, but if this could be included within the distance-bounds matrix in some way this might allow the scores to be more accurate. Alternatively, the scaffolds could be used to align the conformers, or all conformers could be aligned to one or more of the ligands. This was not done as part of this analysis as the alignment will make it less likely that molecules that occupy different parts of the binding site will be identified.

Scaffold 17 has been calculated to have a very low Interaction Score but a relatively high Pharmacophore and Element Scores (Figure 2.27) and does not have a scaffold that matches that of the ligands. One molecule exhibits this scaffold, and the conformers of the molecule are shown in the element, pharmacophore and interaction grids in Figure 2.30. The interaction grid is shown in A. There is a set of hydrophobic interactions that overlaps with hydrophobic interaction counters in the grid which could be the cause of the low Interaction Score. In terms of the element grid, C, and pharmacophore grid, B, there is little overlap because the conformers are all pointing out of the other side, away from the grid. Again, it would perhaps be more useful to align the conformers to the grid or perform constrained docking to enable a more accurate score to be calculated.

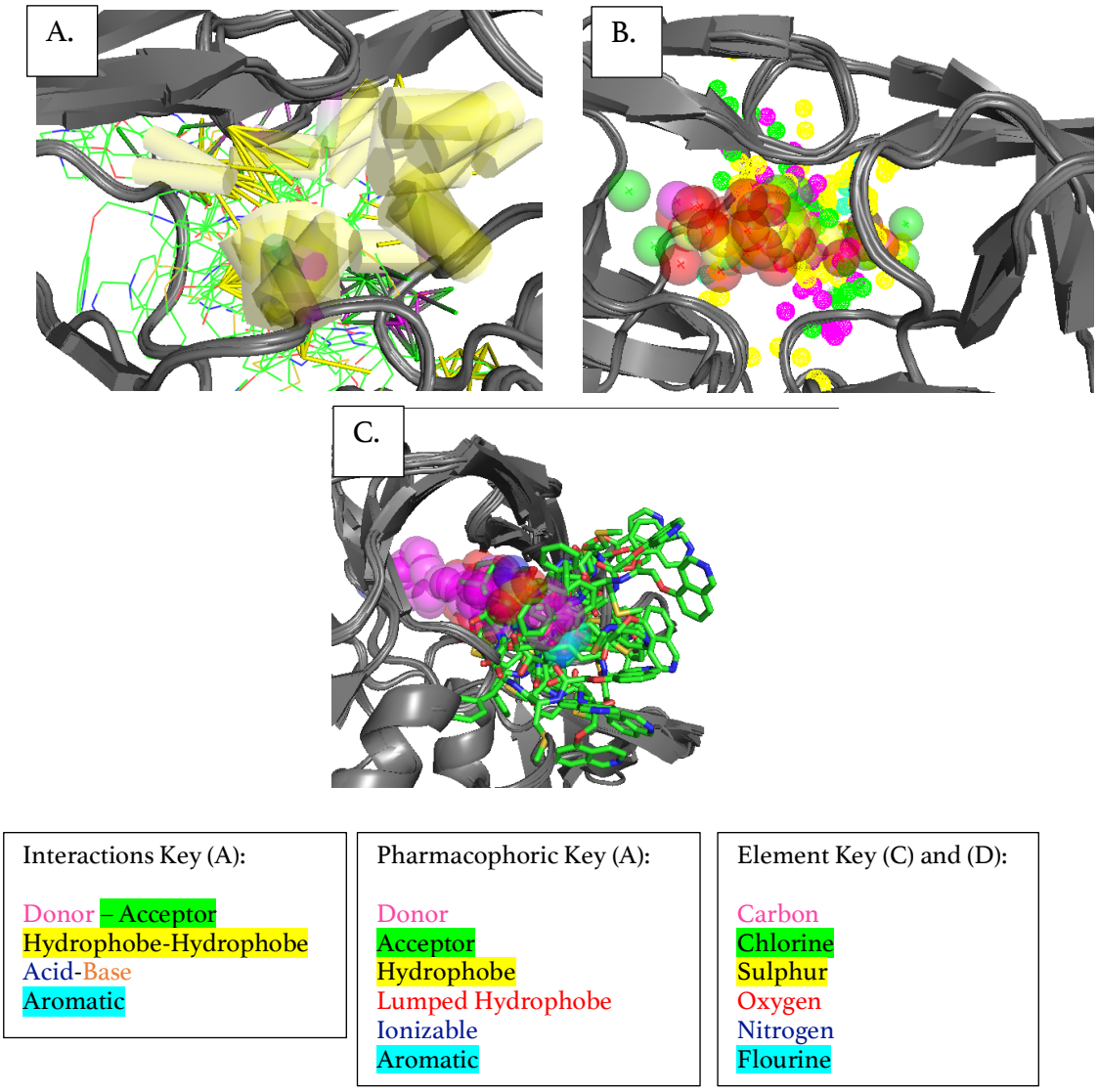



Figure 2.30 Conformers of a molecule that has a different scaffold to those of the ligands used to generate the grid shown in A: the interaction grid, with interaction counters shown transparent and interactions of the conformer opaque, B: the pharmacophore grid shown as transparent spheres with pharmacophoric features shown as opaque spheres, C: the element grid including carbon and D: the element grid without carbon. The size of the sphere or cylinder representing the grids indicates the count, and the colour indicates the type of feature. The molecule exhibits a low Interaction Score due to the overlapping hydrophobic interactions with the grid. However high Element and Pharmacophore Scores are calculated due to a lack of overlap with the grid. Images created using PyMOL (Schrödinger, LLC).

These results indicate that the CRANkS algorithm has potential to be used as a scaffold-hopping technique. Scaffolds that are also found in the set of bound ligands to make the grid are given low Interaction, Pharmacophore and Element Scores. Conversely different scaffolds are given higher Pharmacophore and Element Scores, and the Interaction Score can be used to identify scaffolds that fulfil similar protein-

ligand interaction to those already observed in the structural data and used to build the grid. However, it would perhaps allow the scores to be more accurate if the conformers can be aligned to the binding pocket. The conformers point out in all directions and the ligands all occupy relatively the same area of the binding site with similar orientations. Making use of this prior knowledge to align the conformers should allow more accurate conformer generation in terms of the RMSD with the crystal structure and should allow more accurate CRANkS scores to be calculated.

### *2.3.7 Comparison with Other Dissimilarity Metrics (Set 2)*

To investigate whether the CRANkS scores are representing the molecular novelty adequately, the scores must be compared to other molecular dissimilarity methods. The Element Score has been compared to the Tanimoto coefficient (Section 1.3.1) using two types of chemical fingerprinting shown in Figure 2.31. Morgan fingerprints and 3D Pharmacophore fingerprints were used (Rogers and Hann, 2010; Landrum, 2015, see Section 3.2.5.1). The Tanimoto similarity was calculated between the molecule in question and all the molecules in the grid and then averaged to give the Tanimoto score. This was then taken away from one so that both scores ranked low scores as least novel and high scores as most novel. Both the Element Score and average Tanimoto scores were calculated for both the candidate compounds and the bound ligands used to create the grid to determine whether any scores can distinguish between the molecules in the grid and the molecules not in the grid - molecules used to create the grid are inherently less novel compared to the grid as they were used to create the grid and so should be scored lower.

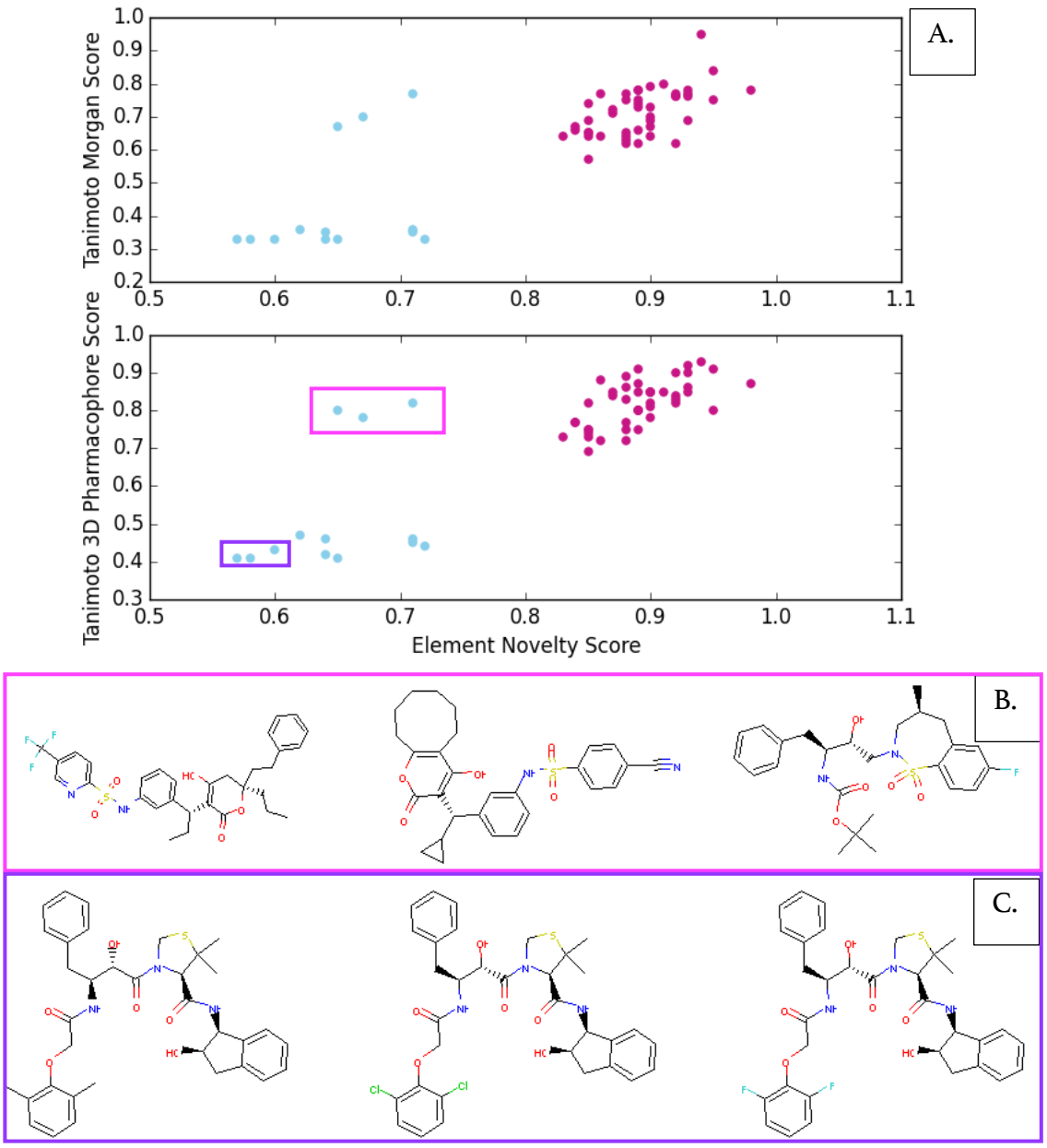



Figure 2.31 A. Comparison of the Element Score and Tanimoto score using two different fingerprints to score molecules used in the grid generation and candidate compounds not used to generate the grid. The molecules used in grid generation are shown in blue and candidate compounds are shown in pink. The Element Score shows clear separation of the two sets of molecules whereas the Tanimoto score does not. B. shows the compounds that were calculated to have a low similarity by the Tanimoto scores compared to the other ligands used to construct the grids (circled in pink in A.). C. shows example compounds used to construct the grids calculated to have a high similarity by the Tanimoto scores – these are clearly similar compounds explaining the high similarity calculated (circled in purple in A.).

The Element Score shows clear differentiation between molecules used in the grid and the candidate compounds - molecules in the grid are scored significantly lower. In contrast for both fingerprinting methods the Tanimoto score does not

differentiate as clearly between the two sets, with three of the molecules in the grid exhibiting scores as high as molecules not used in grid generation. Many of the compounds used in the grids are similar to each other (as shown in Figure 2.31 C) which causes the Tanimoto scores of the majority of the compounds in the grids to be scored low (less novel *i.e.* more similar). The three compounds calculated to have a low similarity by the Tanimoto scores are shown in Figure 2.31 B. These compounds were not similar to any of the other compounds used in the grids. In contrast by using the Element Score these compounds were found to have similar to scores to the other ligands used to construct the grids, by using overlap of atoms within the binding site to calculate the score. The Element Score is therefore offering some new information beyond the conventional molecular dissimilarity score.

### *2.3.8 Bromodomain of Human Bromodomain Containing Protein 1 - BRD1*

Bromodomains exist as an alpha-helical bundle with a central binding pocket that binds acetylated lysine on N terminal histone tails. Bromodomains play a key part in the regulation of the transcription of DNA. Bromodomains have been a target of interest at the Structural Genomics Consortium (SGC) in Oxford owing to their role in wide range of cancer cells (Mertz *et al.,* 2011; Fiskus *et al.,* 2014; Da Costa *et al.*, 2013) and in inflammatory disorders when bound by a small molecule (Wang *et al.,* 2010). However, the exact manner of how this occurs is not well- characterised. Consequently, the SGC have worked to develop chemical probes for a number of bromodomains, including the bromodomain of human bromodomain containing protein 1 (BRD1).

### *2.3.9 Pre-processing of BRD1 Data*

Two sets of data were taken from the SGC Oxford as a test dataset for the CRANkS algorithm for the target BRD1. The first is a set of protein-ligand X-ray crystal structures. BRD1 exists as a dimer in protein-ligand crystal structures, as shown in Figure 2.32, but as a single chain in solution. There were 45 individual protein-ligand structures taken from the SGC Oxford database. However, as these existed as dimers and the ligand could be bound to either chain, each individual chain was treated as a separate structure. These were aligned using the backbone atoms of the proteins and any duplicates or allosteric binders were removed. This left 58 structures to be used to generate the CRANkS grids, listed in Appendix B, Table B.2.5.

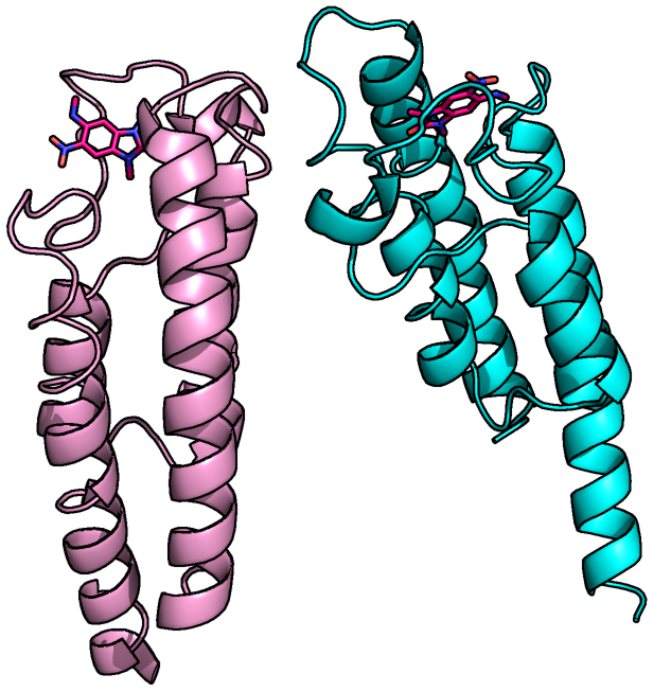


Figure 2.32 X-ray crystal structure for BRD1 for PDB structure 5p07. The two chains shown in pink and cyan only exist as a dimer due to crystal packing. However, the chains have nearly identical conformations and, in this case, both chains have bound the ligand.

A set of 1497 candidate compounds for testing was constructed using assay data from the SGC Oxford. These compounds were tested using three different assays: the amplified luminescent proximity homogeneous assay (AlphaScreen), isothermal titration calorimetry (ITC) and thermal shift assay (TSA). If compounds were tested via multiple assays the most accurate method was taken. A cut-off ligand efficiency of 0.28 was taken to classify actives by AlphaScreen or ITC, based on an $IC_{50}$ of 100 μM (Shultz, 2013; Filippakopoulos *et al.,* 2010). In the case of TSA, a shift of 1°C or greater was taken to classify actives. Therefore compounds with a ligand efficiency of less than 0.28 or a TSA of less than 1°C were classed as inactive. The compounds are not listed in the Appendix due to confidentiality agreements.

### *2.3.10 Discrimination Between Actives and Inactives (BRD1)*

The candidate compounds were ranked using each of the CRANkS novelty scores from least novel to most novel. I hypothesise that the compounds that are less novel in terms of comparison with the grid are more likely to be active, as the compounds are more similar in placement of the binding site, pharmacophoric features or how the compound interacts with the protein. This can be investigated by plotting a receiver-operator-characteristic curve (ROC) which is discussed in Section 3.2.3. The ROC curve for the BRD1 dataset is shown in Figure 2.33. The curves generated using the CRANkS Interaction, Element and Pharmacophore Score are shown. These are compared to using Morgan Fingerprints or Protein-Ligand Interaction Fingerprints (PLIFs) (Roger and Hann, 2010; Bradley *et al.,* 2015; Da *et al.,* 2014). PLIFs were generated either using each protein residue as a bit or using each grid point in the binding pocket as a bit. Notably, the CRANkS Interaction Score outperforms all

other methods with the largest area under the ROC curve and the greatest early enrichment. This indicates that the CRANkS algorithm has potential to discriminate between actives and inactives based on the similarity of compounds to the CRANkS grids.

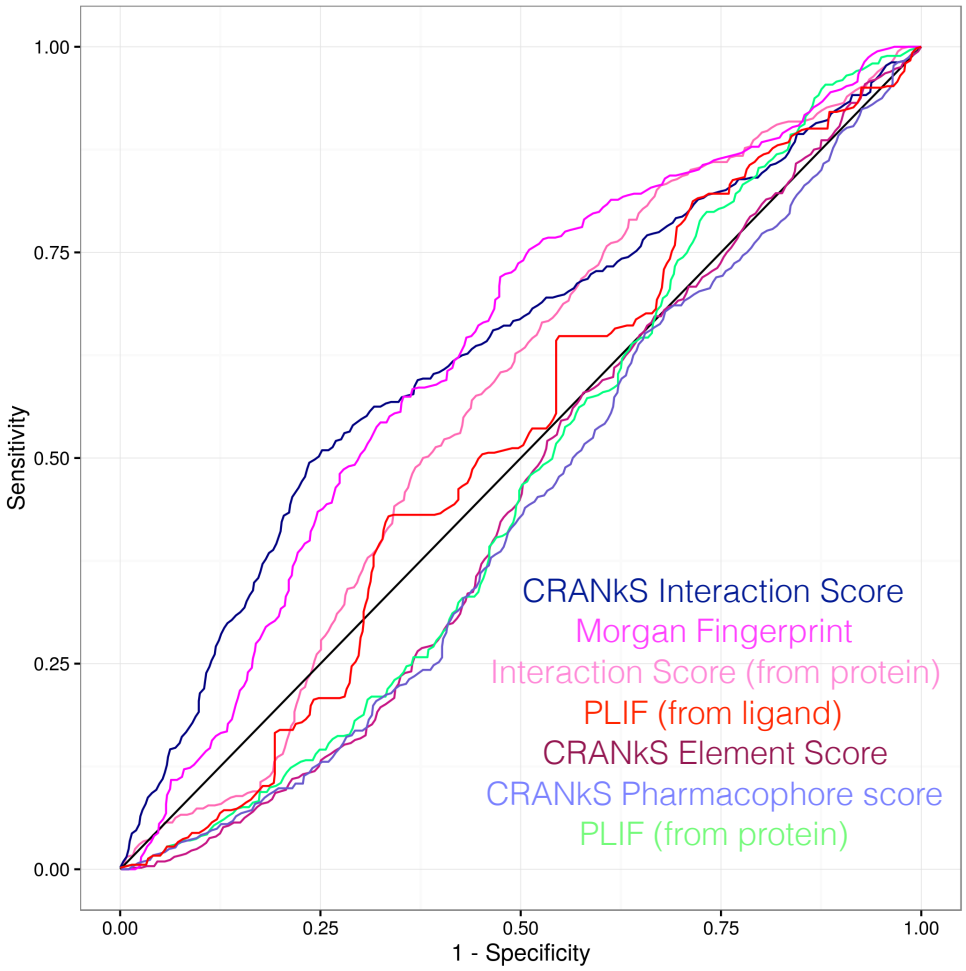



Figure 2.33 ROC curves for the ranking of the BRD1 dataset for the three CRANkS scores, Protein-Ligand Interaction Fingerprints using either each bit as a residue of the protein, or a grid point in the binding site, and Morgan Fingerprint similarity. The CRANkS Interaction Score showed the highest early enrichment and greatest area under the curve.

## 2.4 Further Work

### *2.4.1 Development of CRANkS Algorithm*

#### *2.4.1.1 Additional Molecular Interactions*

Currently the CRANkS algorithm only calculates four types of interactions between pharmacophoric features using RDKit. However, many more molecular interactions are known to be important for protein-ligand binding (Bissantz *et al.*, 2010) and consequently an expanded set of molecular interactions will be implemented in the CRANkS algorithm. There are many different definitions of molecular interactions. Currently, CRANkS defines the four types of interactions using definitions from the Protein-Ligand Interaction Profiler (PLIP) (Salentin *et al.,* 2015). PLIP now calculates eight different types of interactions: hydrogen-bonding, hydrophobic, π-stacking, π-cation interactions, salt bridges, water bridges, halogen bonds and metal complexes.. Additionally, interaction detection is currently the process that takes the longest as part of the CRANkS algorithm, so efforts will be made to speed up this part of the method. PLIP now offers a command-line tool and the addition of using the PLIP command-line tool will allow for the inclusion of water to calculate water-mediated hydrogen bonds, which are known to be important in protein-ligand binding.

#### *2.4.1.2 Conformer Generation*

Previous studies have found the performance of the RDKit conformer generation to be correlated with the number of rotatable bonds, and also generated conformers with an RMSD of less than 2 Å for molecules with 12+ rotatable bonds (Ebejer *et al.*, 2012). However, this does not involve generation of the conformer within the binding site, and aligns the conformer to the crystal structure before calculating the RMSD. In the case of the CRANkS algorithm the conformer is not aligned to the crystal structure, as the calculation needs to reflect whether the two binding poses are close to each other both in terms of the orientations of atoms relative to each other but also in terms of the placement in the binding site. Many conformers generated in this study were found to be approximately 20 Å away from the original crystal structure. This could have a vast effect on the performance of the algorithm and thus more accurate methods of conformer generation should be investigated.

### *2.4.2 Testing of the CRANkS Algorithm*

The work described in this chapter only describes preliminary testing of the algorithm on two targets: HIV-1 protease and BRD1. The CRANkS scores are compared to Morgan Fingerprints and PLIFs but not to any other tools. Therefore, a much larger scale testing of the CRANkS Algorithm was required. Scoring algorithms are known to be target-dependent so a variety of targets should be tested. Many benchmark datasets exist for benchmarking virtual screening tools, for example the Database of Useful Decoys: Enhanced (DUD-E) (Mysinger *et al.*, 2012)

that could be used to test the data and compare the results with other computational algorithms. The CRANkS algorithm should be tested not only in terms of discrimination of actives versus inactives, but also the diversity of scaffolds prioritised as ideally the algorithm should prioritise novel active compounds that ideally facilitate scaffold-hopping.

## 2.5 Conclusions

I have developed the CRANkS algorithm to rank candidate compounds based on how novel the compounds are compared to known binders. The algorithm is interaction-centric and ranks molecules based on how novel the molecules are in terms of chemical structure, placement in the binding site and interactions with the protein. The algorithm is data-driven and constructs grids that describe protein-ligand structural data to score compounds. Initial testing of the algorithm presented here indicates promising results. I have shown the algorithm is able to discriminate between actives and inactives better than when using 2D fingerprint similarity for a HIV-1 protease dataset. More importantly, the algorithm has been shown to have some scaffold-hopping potential when selecting compounds based on the CRANkS scores.

Further work on the algorithm would include more protein-ligand interactions and to include bridging waters. Additionally, the conformer generation step must be investigated to determine whether more accurate conformations can be generated. Finally, a larger-scale testing of the algorithm on multiple datasets is required to investigate the use of the algorithm as a drug discovery tool. A comparison of the tool

to other virtual-screening methods and similarity tools is also required to further investigate the usefulness of the CRANkS algorithm. This work is described in Chapter 3.

# Chapter 3 CRANkS Algorithm II

## 3.1 Introduction

In Chapter 2 I describe the CRANkS algorithm. The algorithm uses structural data of protein-ligand complexes to build grids. These grids are then used to score candidate compounds in terms of how novel the compounds are in comparison with the structural data. Initial results from the algorithm were promising in terms of active versus inactive discrimination and prioritising novel compounds.

In this chapter I examine the benefit of the CRANkS algorithm for use in initial hit-to-lead development. I show the algorithm has comparable ability to leading commercial docking tools in separating actives from inactives. However, the real benefit to the algorithm comes from its ability to prioritise novel active scaffolds maintaining a higher scaffold diversity, and consequently providing better coverage of chemical space.

To investigate the performance of the CRANkS algorithm, it must be tested in terms of both recovering actives and the scaffold diversity of the recovered molecules. It must also be compared to other computational methods. The algorithm was tested on datasets for a variety of targets, as the performance of nearly all computational methods are target-dependent. A portion of these datasets were taken from the well-known benchmarking set The Database of Useful Decoys: Enhanced (DUD-E)

(Mysinger *et al.*, 2012) which allowed comparison of the CRANkS algorithm with commercial docking tools and other methods reported in the literature.

The CRANkS algorithm uses multiple protein-ligand X-ray crystal structures to construct knowledge-based grids which are then used to score candidate compounds. The process for testing the algorithm is illustrated in Figure 3.1. The algorithm was tested on sets of 5, 10, 25 and 50 randomly selected input structures, where possible, for each target. For each test set the experiment was repeated five times, each time using a different set of randomly chosen crystal structures. By using different numbers of structures to build the CRANkS grids, but additionally different randomly chosen subsets, the sensitivity of the CRANkS algorithm to the numbers of structures used, but also the effect of which structures are used to build the CRANkS grids, was tested.

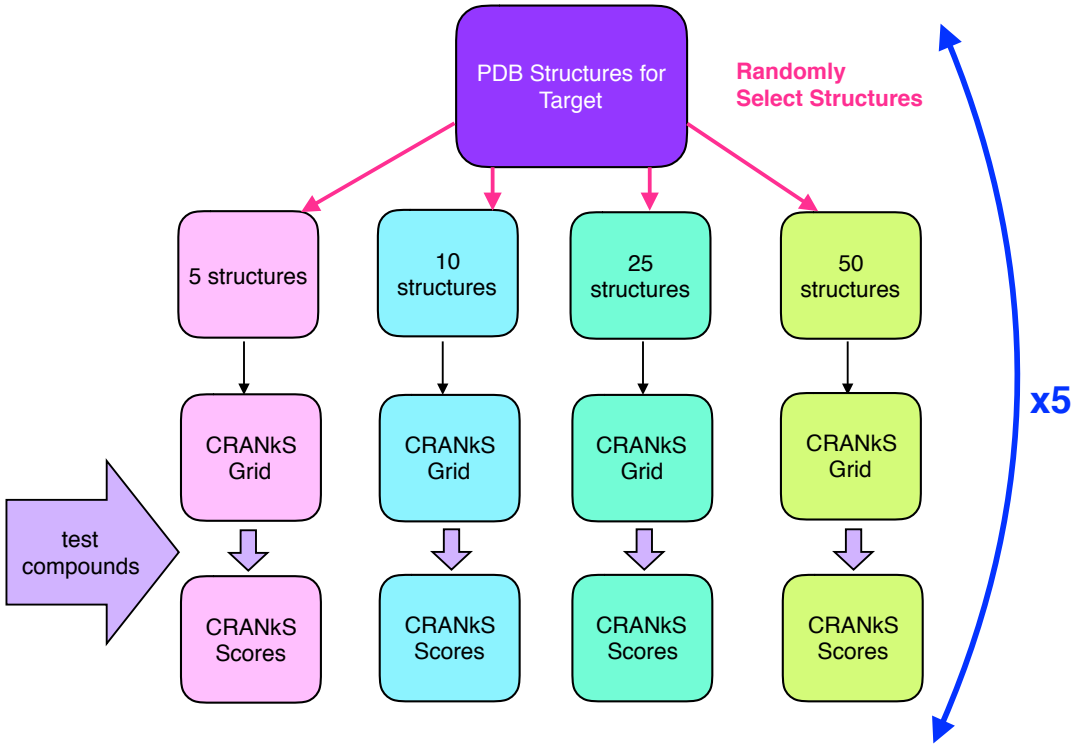



Figure 3.1. Depiction of how the CRANkS algorithm is tested using benchmark sets. For each target with a set of protein-ligand structures, sets of 5, 10, 25 or 50 structures are randomly selected where possible and used to generate CRANkS grids to score the test compounds. This is repeated 5 times for each set of numbers of structures.

The three CRANkS scores were calculated for each of the test compounds for each set of constructed grids. The compounds were then ranked based on these scores from least novel to most novel. Compounds that overlap better with the grids are assumed to be more similar to the bound ligands, in terms of how the ligands interact with protein and thus more likely to be active. This ranking from least novel to most novel is therefore used to calculate how well the algorithm discriminates between active and inactive compounds using a variety of metrics to look at the enrichment and ranking. These metrics allow the ranking ability of CRANkS to be compared to those reported for popular commercial docking tools. To investigate the recovery of scaffolds in this ranking, the scaffold associated with each test compound was calculated. Again, a variety of metrics are used to look at the enrichment of unique scaffolds earlier in the ranking. These metrics allow comparison to well-known fingerprinting tools which are used regularly to calculate molecular similarity. These are discussed in the Methods section (see Section 3.2).

Results for the discrimination between actives and inactives were promising with the CRANkS algorithm outperforming 3 out of 4 commercial docking tools using the BEDROC metric for the majority of targets (Section 3.3.1). However, using 2D similarity to rank compounds usually performed better in terms of separating actives from inactives. The scaffold-hopping potential was also investigated, and the CRANkS scores were found to prioritise more unique scaffolds earlier in the ranking than fingerprinting scoring methods by both scaffold diversity metrics (Section 3.3.2). Finally, multi-objective optimisation was used to select compounds to test how the algorithm could be used as part of a drug discovery campaign (Section 3.3.2.2). This

was found to select compounds with both a high enrichment of scaffolds and active scaffolds when compared to a random selection of compounds.

## 3.2 Methods

### *3.2.1 Improvements to the CRANkS Algorithm*

#### *3.2.2.1 Conformer Generation*

The conformer generation tool used in CRANkS was updated in two parts compared to the method described in Chapter 2. First, the 3D Matched Molecular Pairs part of the algorithm was updated. The method described in Chapter 2 involves breaking each candidate compound into blocks and matching those blocks with the constituent blocks of the bound ligands. However, this is time-consuming as each compound must be broken multiple times and then a long list of ligand building blocks must be searched multiple times. A more time-efficient solution is to use the Maximum Common Substructure (MCS) detection feature as implemented in RDKit (Landrum, 2015). By using the constraint of only matching complete rings, this is an effective way to quickly find the biggest overlap between the candidate compounds and each of the ligands. This was calculated using RDKit. Conformers are generated with the maximum common substructure between the compound and any of the crystallographic ligands constrained to the coordinates of the crystal structure. A comparison of the two methods used on an example compound is shown in Figure 3.2.

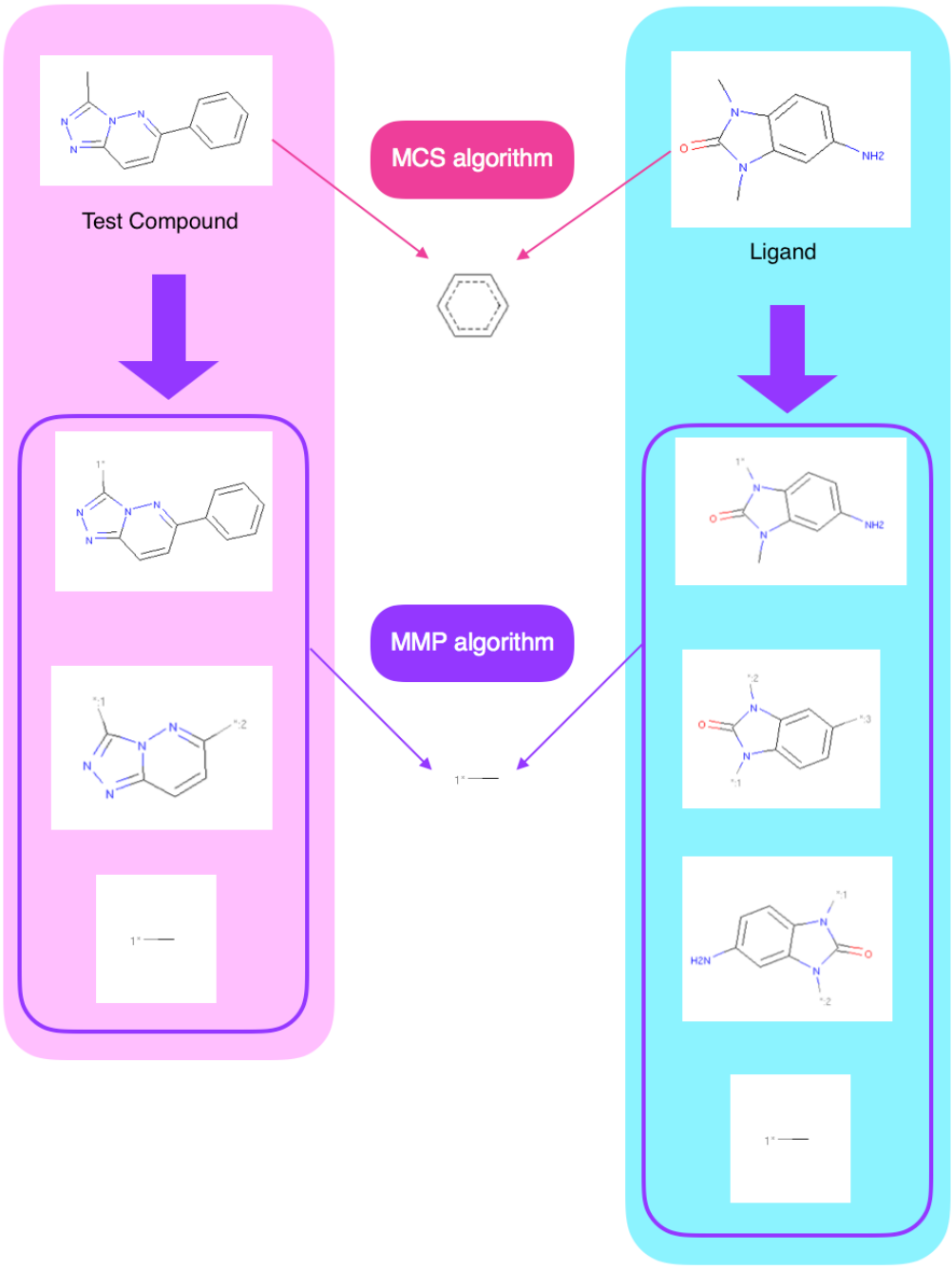



Figure 3.2. Differences in the original MMP algorithm and new MCS algorithm. Using the new algorithm, the test compound would be placed in the binding site using the aromatic benzene ring common to both compounds. In contrast the MMP algorithm only finds a carbon atom common between the two and that is what will be used to place the compound at random.

To test the effect of changing this part of the algorithm, each ligand in the BRD1 dataset (described in Chapter 2) was used as a candidate compound and 100 conformers were generated. All other ligands excluding the ligand in question could be used to place the molecule in the binding site using 3D MMP or the MCS version of the algorithm. The RMSD between the original crystal structure of the ligand and each of the conformers was then calculated. The results are shown in Figure 3.3. For the conformers that fall within 10 Å of the crystal structure, the MCS algorithm generally outperforms the original implementation. It should be noted that for conformers that are very distant to the original crystal structure, approximately 20 Å, the MCS algorithm performs slightly worse than the 3D MMP version, but it is only a marginal difference. The 3D MMP version of the algorithm fails much more frequently when it cannot find a match.

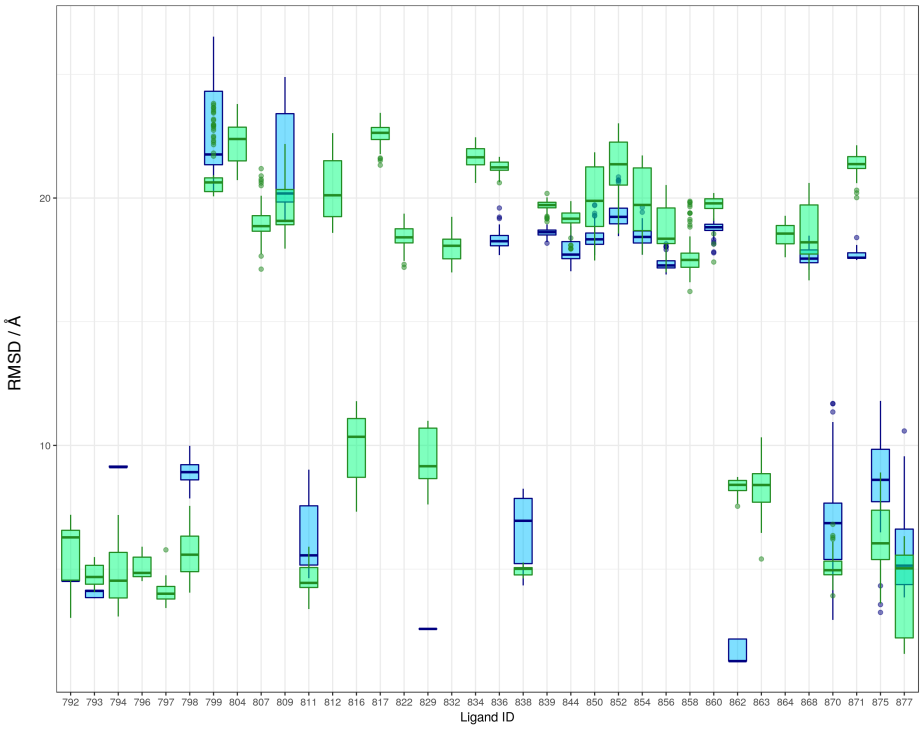



Figure 3.3. The RMSD between the original crystal structure and 100 conformers generated using the CRANkS algorithm. The original 3D MMP method is shown in blue and the new MCS version of the algorithm is shown in green. Overall the MCS version fails less often and generates conformers closer to the crystallographic structure if the ligands are closer than 10 Å.

The method was changed from the original 3D MMP algorithm as described in Chapter 2, to a faster slightly more accurate implementation using MCS. However, it is still clear that conformer generation is likely to be an issue in this algorithm as conformers are still consistently more than 1.0 Å away from an original crystal structure.

Secondly, in the original implementation of the algorithm (Chapter 2), conformers would be produced until either 100 conformers had been generated or 5000 conformations had been rejected due to a steric clash with the protein. However, this had not been tested - the number of conformations rejected could have a big impact on computational time if the algorithm was continuing unnecessarily. Figure 3.4 shows the computational time for the conformation generation part of the algorithm for four different targets with respect to the maximum number of rejected conformers. It is clear that although the average computational time for a run of the algorithm remains approximately equal, it is the outliers that cause a significant increase in computational time, as the number of rejected conformers allowed increases. To investigate the effect on performance, Figure 3.5 shows the Area Under the Curve (described in Section 3.2.3.1) for four benchmark datasets of different targets. There is no clear detriment to performance, as measured by AUC, for reducing the maximum number of rejected conformations. As a result, the algorithm was modified so that it would finish when either 100 accepted conformations or 500 rejected conformations were generated.

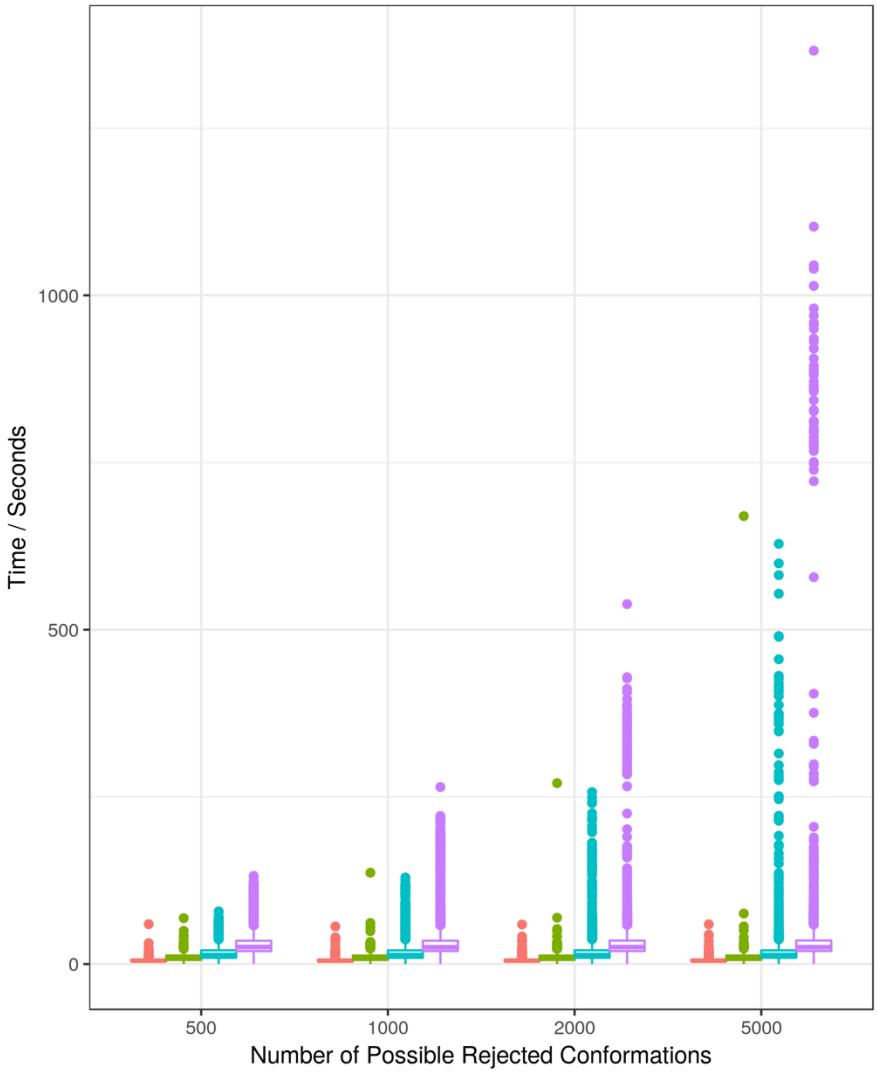



Figure 3.4. The time taken for the CRANkS algorithm with respect to the number of rejected conformations allowed before exiting. This is shown for benchmark datasets for four different targets: BRD1 in orange, DHFR in green, HSP90A in blue and PYGM in purple. Although the average time remains similar across the number of possible rejected conformations, the outliers show a vast increase in computational time.

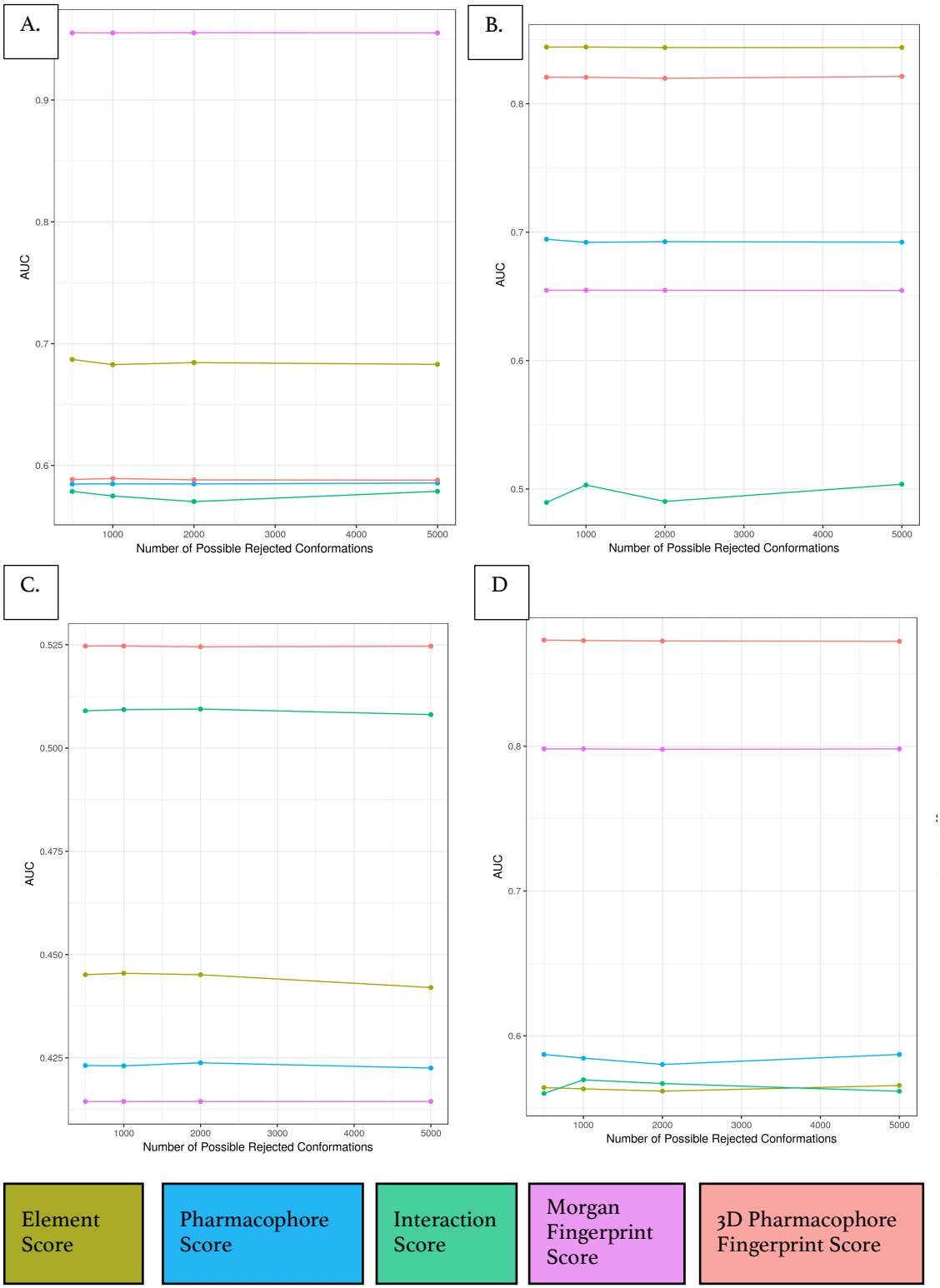



Figure 3.5. The effect of varying the maximum number rejected conformations on the Area Under the Curve is shown for five different scoring methods: the CRANkS scores: Element Score in yellow, Interaction Score in green, Pharmacophore Score in blue and the Morgan Fingerprint Score in pink and the 3D Pharmacophore Fingerprint Score in orange. This shown for four different targets, in A: PYGM, B: HSP90A, C: BRD1 and D: DHFR. The number of possible rejected conformations shows little effect on the performance of the scores in terms of AUC.

### *3.2.2.2 Calculation of Interactions*

The original implementation of the CRANkS algorithm (described in Chapter 2), calculated interactions by calculating pharmacophoric features on both the protein and ligand and using definitions from the Protein- Ligand Interaction Profiler, PLIP (Salentin *et al.,* 2015), to calculate whether an interaction was present. This original implementation only included four types of interactions. However, many more molecular interactions are known to be important for protein-ligand binding (Bissantz *et al.,* 2010). Additionally, water was neglected as part of the original implementation of the algorithm. Water meditated hydrogen bonding, *i.e.* water bridges occur frequently in protein- ligand complexes and can have an important effect on binding (see Section 1.4.3).

PLIP now calculates eight different types of interactions: hydrogen-bonding hydrophobic, π-stacking, π-cation interactions, salt bridges, water bridges, halogen bonds and metal complexes (Salentin *et al.,* 2015). A command-line version of the tool was also developed. As this method has been rigorously tested, the CRANkS algorithm was modified to make use of the PLIP command-line tool to calculate interactions. This expands the types of interactions available, allowing water in the original protein-ligand structure to be used to calculate water-mediated hydrogen bonds.

### *3.2.2 Datasets*

Datasets were required that contain both active and inactive compounds to test the CRANkS algorithm. Many benchmarking datasets exist to test docking tools and other virtual screening methods including DUD (Huang *et al.*, 2006), DUD-E (Mysinger *et al.,* 2012), and MUV (Rohrer and Baumann, 2009). By using a benchmarking set, the results for the CRANkS algorithm can be compared to commercial docking tools and other algorithms tested on the same set. In this case the benchmarking set used was The Database of Useful Decoys Enhanced (DUD-E) (Mysinger *et al.*, 2012). It consists of sets of actives for each target with a "measured affinity supported by a literature reference" and a generated set of decoys. These sets are for 102 targets that fit into 8 broad target classes. One target was chosen from each drug target class. Targets were selected with as many PDB structures as possible whilst minimising the number of test molecules. The chosen datasets are summarised in Table 3.1.

| Target Class | Database | Target | Acronym | Number of Actives | Number of Decoys | Number of PDB Structures (DUD-E) | Number of PDB Structures After Inspection |
|---|---|---|---|---|---|---|---|
| GPCR | DUD-E | Beta 2 adrenergic receptor | ADRB2 | 231 | 15000 | 7 | 7 |
| Nuclear Receptor | DUD-E | Androgen Receptor | ANDR | 269 | 14350 | 50 | 21 |
| Protease | DUD-E | Beta-secretase 1 | BACE1 | 283 | 18100 | 129 | 128 |
| Ion Channel | DUD-E | Glutamate Ionotropic Receptor AMPA Type Subunit 2 | GRIA2 | 158 | 11845 | 80 | 30 |
| Miscellaneous | DUD-E | Heat Shock Protein 90-alpha | HSP90A | 88 | 4850 | 76 | 76 |
| Other Enzymes | DUD-E | Glycogen Phosphorylase, muscle form | PYGM | 77 | 3950 | 125 | 87 |
| Kinase | Literature | Cyclin-dependent kinase 2 | CDK2 | 161 | 3304 | 157 | 153 |
| Bromodomain | SGC Oxford | Bromodomain-containing protein 1 | BRD1 | 547 | 806 | - | 46 |
| Enzyme | Literature | Dihydrofolate reductase | DHFR | 50 | 706 | 75 | 65 |

Table 3.1. Benchmark datasets used for testing the CRANkS algorithm. Soluble proteins are coloured in blue and membrane proteins in green. Targets using grids with only 5 compounds are coloured in dark purple, 5 and 10 compounds in light purple, 5, 10 and 25 in in light pink and 5, 10, 25 and 50 in dark pink.

One issue with using the DUD-E dataset is that although matched by physiochemical properties, the putative decoys are inherently chemically dissimilar to the actives in terms of chemotype, due to the way the decoys are generated. This is unlike an actual drug campaign when tested molecules would tend to be much more chemically similar in 2D. Therefore, more challenging datasets were gathered to test how the algorithm would work with experimentally proven inactives. These datasets were collated for Dihydrofolate reductase (DHFR), Bromodomain-containing

protein 1 (BRD1) and for Cyclin-dependent kinase 2 (CDK2). The additional datasets are discussed below.

The dataset for CDK2 and the dataset for DHFR were taken from an investigation of Feature-Map Vectors (Landrum *et al.*, 2006). The CDK2 dataset is a subset of a 16000-compound dataset from a library design investigation that had been synthesised and screened for a CDK2 antagonists project (Bradley *et al.*, 2003; Sielecki *et al.*, 2000; Knockaert *et al.,* 2002). The subset was chosen to include molecules that were specifically synthesised to target CDK2 or were acquired to specifically target CDK2. Consequently, the molecules are rich in features known to be important for activity against CDK2 and represent a more demanding test set for an algorithm to separate actives and inactives. Compounds from screening libraries were discarded and molecules described as 'Moderately Active' were also discarded. The dataset for DHFR is a set of 756 compounds with a range of $IC_{50}$ values (Sutherland *et al.*, 2003). These were converted into discrete active and inactive labels, with active molecules requiring an $IC_{50}$ of better than 0.04 μM and all other molecules labelled inactive (Landrum *et al.*, 2006). The BRD1 dataset was taken from screening data for the Structural Genomics Consortium in Oxford and is described in Chapter 2 Section 2.3.3.

For targets from the DUD-E database and for the CDK2 dataset protein-ligand structures were taken directly from the DUD-E website (stored as "*pdb_analyze.txt*"). These protein-ligand structures were collated as part of DUD-E by selecting structures from the PDB using the UniProt code of the target (Appendix Table B.3.1 – B.3.7). The PDB structures were aligned in PyMOL using the *align* function. The

binding site was determined by visual inspection of the PDB structure chosen to be docked to in the DUD-E dataset. Once aligned any PDB structures where the ligand is not bound within the binding site, where the ligand is a repeat of another ligand already within the set and in the same conformation, or where the protein structure is extremely different are removed. Tables B.3.1 – B.3.7 in Appendix B detail the PDB structures within each set including the reasons for rejection for structures removed from the set for testing with the CRANkS algorithm. For BRD1, the PDB structures used are shown in Appendix B in Table B.2.5. For target DHFR PDB structures were collated from the PDB using the UniProt code of the target and the structures used are listed in Table B.3.8 in Appendix B.

The RMSD between each of the PDB structures for the whole dataset of structures for each target using the whole atom structure are shown in Figures A.3.1 – A.3.8 in Appendix A. The randomly chosen PDB structures to be used in the construction of each of the CRANkS grids for each target are shown in Tables B.3.9 – B.3.137 in Appendix B, including the residue code for the ligand and the resolution of each structure.

### *3.2.3 Metrics for Discrimination Between Actives and Inactives*

To test the ability of the CRANkS algorithm to discriminate between active and inactives, a number of metrics can be calculated from the order of the compounds based on each CRANkS score. All of the metrics were calculated using the R package '*enrichvs*'. (H. Yabucchi, 2011). I treat the output from the CRANkS algorithm as a binary classification model where compounds are either labelled as active (*a*) or

inactive (*i*). The number of compounds that are active for each target are known and are shown in Table 3.1.

This binary classification allows there to be four possible outcomes. The algorithm can classify a compound *a* when the compound is actually active, which would be a true positive. A true negative is therefore when an inactive compound is correctly classified as *i*. If a compound that is actually active is classified by the algorithm as *i,* this outcome would be a false negative. Conversely, if a compound that is actually inactive is classified as *a,* the outcome is called a false positive. The four outcomes make up a contingency table or confusion matrix, which is shown in Figure 3.6. The true positive rate (TPR) and false positive rate (FPR) are also shown, which are calculated using the instances of the matrix.

| | | True Condition | |
|---|---|---|---|
| | Total Population | Condition Active | Condition Inactive |
| Predicted Condition | Predicted Condition Active | True Positive | False Positive, Type I Error |
| | Predicted Condition Inactive | False Negative, Type II Error | True Negative |
| | | TPR = $\frac{\Sigma \text{ True positive}}{\Sigma \text{ Condition positive}}$ | FPR = $\frac{\Sigma \text{ False positive}}{\Sigma \text{ Condition negative}}$ |

Figure 3.6. Confusion matrix for treating the CRANkS algorithm as a binary classifier.

From the confusion matrix, the receiver-operator-characteristic (ROC) curve can be drawn. This illustrates the diagnostic ability of the algorithm as a dependence of the discrimination threshold, $T$. The ROC curve plots the true positive rate (TPR) as a function of the false positive rate (FPR). An example of the distribution of the Morgan Fingerprint Score (Section 3.2.5.1) for active and inactive compounds of DHFR is shown in Figure 3.7. The different classifications of the confusion matrix are coloured on the distribution for a given $T$. The corresponding ROC curve is also shown. The overlap between the two distributions dictates the shape of the curve.

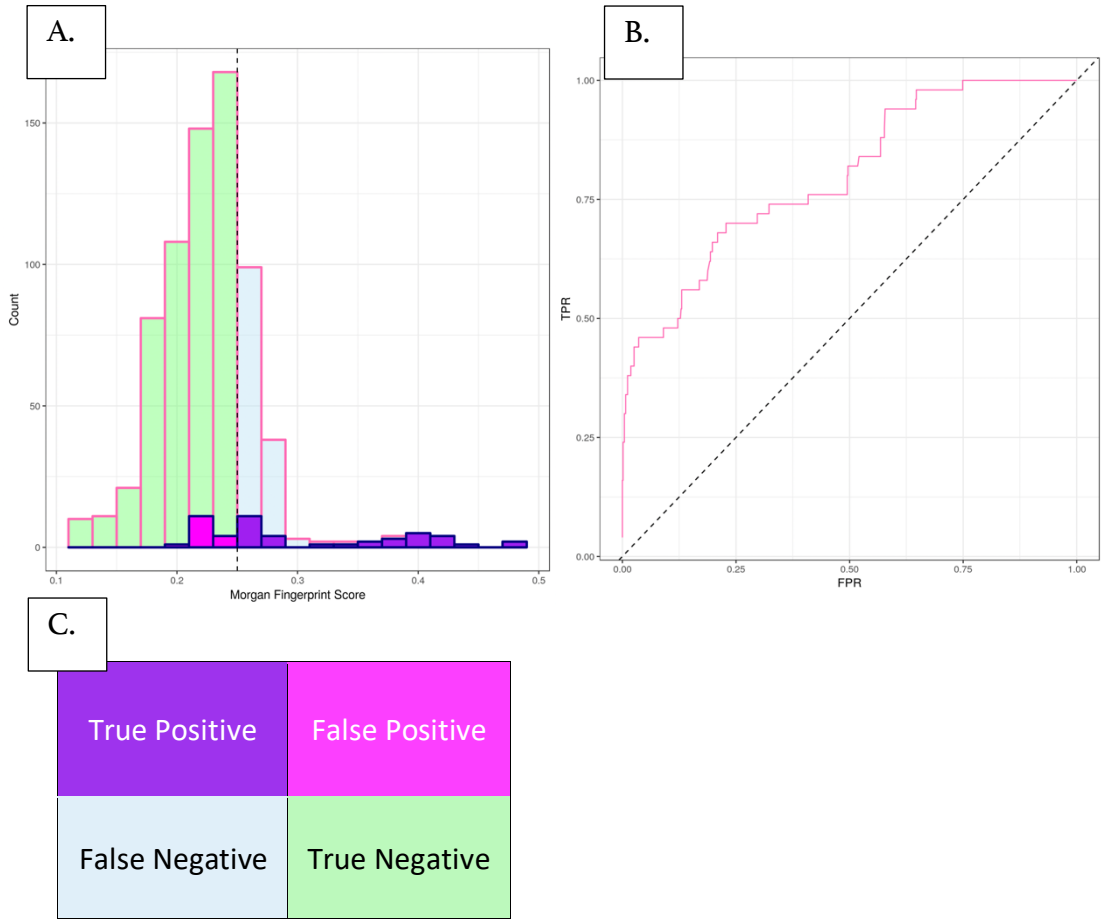



Figure 3.7. The results of using the Morgan Fingerprint Score to rank candidate compounds for DHFR. The distribution of scores for the compounds are shown in A; the distribution of the active compounds is outlined in navy blue and the inactive distribution is outline in pink. A token threshold of 0.25 is shown by the black dotted line. The instances of the confusion matrix are shown in C and the distributions in A are coloured accordingly. B shows the corresponding ROC curve for the data, illustrating how the overlap of the distributions affects the shape of the ROC curve.

The TPR is also known as the sensitivity and can be interpreted as the probability of detection. It defines how many of the active molecules will be correctly labelled as *a*. The FPR is equal to one minus the specificity. It defines the number of inactive molecules that will be incorrectly labelled as *a*. Thus, the ROC space represents the trade-off between the advantage of finding true positives and the detriment of incorrectly predicting false positives as defined by the TPR on the *y*-axis and the FPR on the *x*-axis. Each point in the ROC space represents the prediction of a compound and consequently a case in the confusion matrix.

Within the ROC space, the line of discrimination represents the result of a random guess and is the diagonal that divides the space (shown in Figure 3.7 C as a dashed line). Points that lie above the line of discrimination indicate a classification that is better than random. Perfect classification corresponds to no false positives and thus 100% specificity and 100% sensitivity. It exists in ROC space as the point (0,1).

In this case, the binary classification is based on one of the CRANkS scores, $S$, which is continuous. Using a given threshold $T$, each compound is classified as *a* if $S < T$ and *i* otherwise. If the algorithm used is the Morgan Fingerprint Score or 3D Pharmacophore Fingerprint Score, each compound is classified as *a* if $S > T$ and *i* otherwise. The ROC curve therefore parametrically plots TPR($T$) and FPR($T$) while varying the threshold parameter $T$. If $X$ follows a probability density $f_1(x)$ when $X$ is classified as *a* and $f_0(x)$ if it is classified as *i,* the TPR and FPR can be written as:

$$\mathrm{TPR}(T) = \int_T^{\infty} f_1(x)\, dx \tag{3.1}$$

$$\mathrm{FPR}(T) = \int_T^{\infty} f_0(x)\, dx \tag{3.2}$$

### *3.2.3.1 Area Under the ROC Curve*

The area under the ROC curve (AUC) is a metric that is commonly used to quantify the predictive performance of a binary classifier. In this case, it is the probability that a randomly picked active will be scored lower than a randomly chosen inactive in the case of the CRANkS algorithm, or higher in the case of the fingerprinting scores. It can therefore be written as Equation 3.3 where $X_1$ and $X_0$ is the score of an active and inactive respectively. All other variables are defined previously.

$$\mathrm{AUC} = \int_{\infty}^{-\infty} \mathrm{TPR}(T)\big(-\mathrm{FPR}'(T)\big)\, dT = P(X_1 > X_0) \tag{3.3}$$

The AUC is therefore bounded by 0 and 1, where an area of 1 would correspond to perfect classification, and 0.5 would be random classification. It is also independent of the ratio of actives to inactives allowing direct comparison across targets. However, there has been criticism of using this metric as a method of testing virtual screening methods (Truchon and Bayly, 2007). One limitation is its insensitivity to early recognition. In a drug discovery campaign, virtual screening algorithms will be used to rank and select top scoring compounds for synthesis based on the score. Therefore, the performance of a classifier at the start of the ordered list of compounds can be considered as more useful metric.

### *3.2.3.2 Enrichment Factor*

The Enrichment Factor (EF), in contrast to the AUC, only considers initial enrichment. It is a measure of how many actives are within a certain initial fraction of the ranked compounds, in comparison to a random distribution of compounds. The enrichment factor is defined as the ratio between the percentage of active compounds in the selected subset and the percentage in the entire database. In this analysis, the EF at 1% was considered as this allowed comparison to values from the literature for other algorithms. An EF of greater than 1 implies successful predictions. A caveat of using this metric is that it is dependent on the ratio of the number of actives to the number of inactives, and so cannot be compared across targets.

### *3.2.3.3 Boltzmann-Enhanced Discrimination of Receiver Operating Characteristic*

The Boltzmann-Enhanced Discrimination of Receiver Operating Characteristic (BEDROC) (Truchon and Bayly, 2007) is a metric that was derived to solve the limitation of insensitivity to early recognition of the AUC. BEDROC makes use of the robust initial enhancement (RIE) (Sheridan *et al.,* 2001) to combine a continuously decreasing exponential weight as a function of rank with the statistical significance from ROC. BEDROC is bound between 0 and 1 and is independent of the ratio of actives to inactives allowing comparison across targets.

The RIE was developed to be less susceptible to extreme differences than EF when there are only a small number of actives. It can be written as:

$$\mathrm{RIE} = \frac{{}^{1}/_{n}\sum_{i=1}^{n} e^{-\alpha X_i}}{{}^{1}/_{n}\left(\frac{1-e^{-\alpha}}{e^{\alpha}/_{N}-1}\right)} \tag{3.5}$$

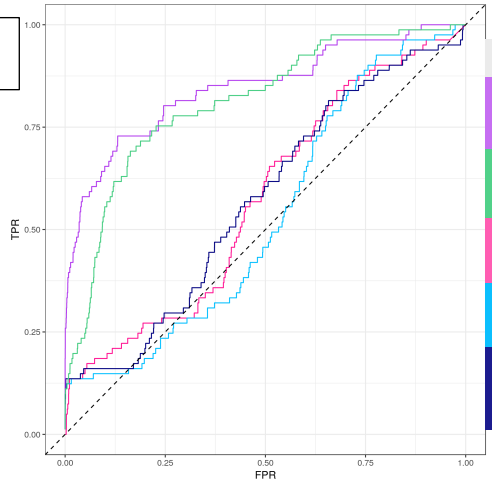
where *n* is the number of actives, *N* is the total number of compounds in the ordered list and α is a constant that is chosen to determine the exponential weight. The BEDROC metric can then be defined as:

$$\mathrm{BEDROC} = \mathrm{RIE} \times \frac{{}^{1}/_{N}\sinh\left({}^{\alpha}/_{2}\right)}{\cosh\left({}^{\alpha}/_{2}\right)-\cosh\left({}^{\alpha}/_{2}-\alpha\times{}^{n}/_{N}\right)} + \frac{1}{1-e^{\propto\left(\frac{N-n}{N}\right)}} \tag{3.6}$$

which is effectively a weighted function of the AUC. In this analysis α = 80.5 was used to allow comparison to values for algorithms found in the literature. This corresponds to the top-ranked 2% of the molecules accounting for 80% of the score. An example comparison of ROC curves for the target GRIA2 (run 5 of 25 structures) is shown in Figure 3.8 for various scoring methods. The corresponding AUC, BEDROC and $EF_{1\%}$ values are shown (Figure 3.8 B). It is clear that although the AUC calculated for the Morgan Fingerprint Score and 3D Pharmacophore Fingerprint Score are similar, the BEDROC values are not, due to differences in the initial enrichment. Equally, although the AUC for the CRANkS Pharmacophore Score is significantly lower than for the 3D Pharmacophore Fingerprint Score, the BEDROC values are similar due to similarities in early enrichment. Conversely, the CRANkS Interaction Score and CRANkS Pharmacophore Score have very similar AUC values,

but the BEDROC value for the Interaction Score is significantly lower due to poorer initial enrichment (Figure 3.8 B).

A.




B.

| Scoring Method | AUC | BEDROC | $EF_{1\%}$ |
|---|---|---|---|
| Morgan Fingerprint Score | 0.85 | 0.48 | 34.5 |
| 3D Pharmacophore Fingerprint Score | 0.81 | 0.22 | 13.6 |
| CRANkS Interaction Score | 0.57 | 0.12 | 9.9 |
| CRANkS Element Score | 0.53 | 0.18 | 12.3 |
| CRANkS Pharmacophore Score | 0.57 | 0.19 | 13.6 |

Figure 3.8. ROC curves (A) and corresponding AUC, BEDROC and $EF_{1\%}$ metrics (B) for a run of 25 structures used in the CRANkS gird for target GRIA2.

## *3.2.4 Metrics for Calculating Scaffold-Hopping Potential*

A scaffold can be defined as a molecular framework to which functional groups are linked (Hu *et al.,* 2016). Scaffolds can be used for two purposes: to identify frameworks that are enriched for activity, but also to find scaffolds that are structurally different but are also enriched for activity *i.e.* scaffold-hopping (Schneider *et al.*, 1999). Scaffold-hopping is of particular importance in drug discovery campaigns, not only for so-called patent busting, but also for a greater exploration of activity space in order to identify the most amenable candidate compound to take forward as a lead.

To show the potential of the CRANkS algorithm to facilitate scaffold-hopping, a scaffold is calculated for each candidate compound. These scaffolds are not unique

and different candidate compounds may have the same scaffold. Metrics were then derived to determine the scaffold diversity of top-ranked compounds by the CRANkS algorithm but also alternative fingerprint methods.

### *3.2.4.1 Scaffold Calculation*

The most commonly used scaffold representation is the Bemis-Murcko Scaffold framework (Bemis and Murko *et al.,* 1996). The molecule is divided into R-groups, linkers and ring systems. The scaffold is created by removing R-groups but retaining rings and the linkers between them. Limitations of this method include that a series of compounds in which rings are added will be treated as having different scaffolds, when in fact it would more useful for the series to be treated as a single scaffold (Hu *et al.*, 2016). Additionally, molecules with no rings will not be given a scaffold. The Bemis-Murcko scaffolds can be further generalised into cyclic skeletons by removing all heteroatoms and replacing them with carbon (Xu and Johnson, 2001). However, this does not seem applicable to this work as removing all heteroatoms is unlikely to yield a scaffold that is representative of activity. In the case of both Bemis-Murcko scaffolds and cyclic skeletons for the DUD-E datasets, nearly all molecules corresponded to a unique scaffold. Thus, it would be nearly impossible to look at scaffold diversity using these definitions as there is no overlap in the molecules in terms of scaffolds.

The method used to derive formal scaffolds in this analysis is Scaffold Hunter (Wetzel *et al.*, 2009; Schäfer *et al.*, 2017). Rather than only define a singular scaffold for the molecule, a scaffold hierarchy is developed by iterative pruning of scaffolds.

Rings are iteratively removed generated a lineage of scaffolds for one compound, with each iteration generating a scaffold of a different level. This results in a scaffold tree and means that by using a certain level of scaffold, multiple molecules in the datasets used here are identified to have the same scaffold. The algorithm creates scaffolds by first removing all side-chains but preserving double bonds attached to a ring. The scaffold is then trimmed by a set of deterministic rules, removing single rings which results in the lowest level scaffold with a single ring (Schuffenhauer *et al.,* 2007). Each iterative step results in a scaffold which allows a scaffold tree to be generated as molecules often share a common scaffold. In this analysis the scaffold used to describe each molecule was the lowest level scaffold *i.e.* that with a single ring. An example of the different scaffolds generated for two candidate compounds is shown in Figure 3.9. The candidate compounds are from the DHFR dataset and could easily be two compounds from the same series. The Bemis-Murcko scaffolds and cyclic skeletons are clearly different for the two scaffolds. However, the Level 2 and Level 1 scaffolds generated by Scaffold Hunter are the same. This indicates how this hierarchical scaffold generation allows for the compounds in the dataset to have identical scaffolds when calculated using Scaffold Hunter but not when using Bemis-Murcko scaffolds. In this work Level 1 scaffolds were calculated by Scaffold Hunter for each of the candidate compounds for each target.

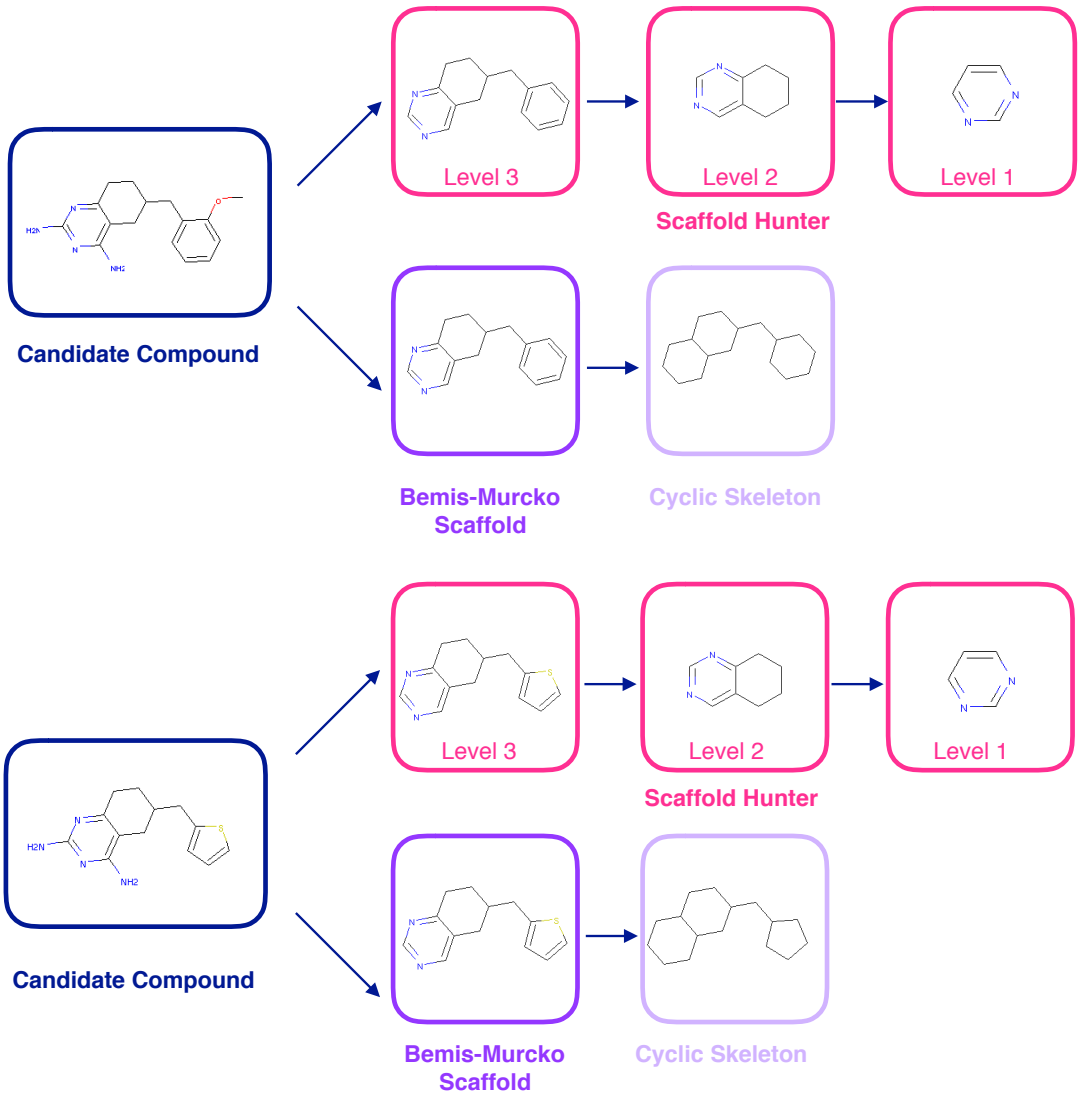



Figure 3.9. Calculated scaffolds for two candidate compounds from the benchmark dataset for DHFR. Using Bemis-Murcko scaffolds or cyclic skeletons yields different scaffolds for the candidate compounds. However, using Scaffold Hunter, the Level 3 scaffolds are different, but the Level 2 and 1 scaffolds are identical.

### *3.2.4.2 Metrics for Scaffold-Hopping Potential*

Two metrics were used to determine the scaffold-hopping potential of algorithms. The first is the number of molecules required to achieve 50% of the scaffolds. This can be used as an overview of the effect of ranking on scaffold diversity. The lower the number of molecules required, the higher the number of unique scaffolds earlier in the ranking, the greater the scaffold diversity, and the greater the propensity for scaffold-hopping. An additional metric was used to give information on the early enrichment of scaffolds. This is the number of unique scaffolds in the top-ranked 100 molecules. In this case, the higher the calculated value, the higher the number of unique scaffolds ranked earlier, the greater the scaffold diversity and the greater the propensity for scaffold- hopping.

In this analysis, the metrics were used considering not only the active scaffolds but also considering the scaffolds of both active and inactive compounds. Considering all the scaffolds is more realistic in comparison to an actual drug campaign when the activity would be unknown. Therefore, the calculated diversity is more like the actual diversity that would be obtained if the algorithm was used. However, there will inherently be an effect on the discrimination of actives with respect to inactives if active molecules have only a few scaffolds.

### *3.2.5 Alternative Methods*

#### *3.2.5.1 Fingerprint Methods*

Two scoring methods were used as a direct comparison to the performance the CRANkS algorithm: the Morgan Fingerprint Score and the 3D Pharmacophore Fingerprint Score. The Morgan fingerprint (using a radius of 2) and 3D Pharmacophore fingerprint were used respectively to describe each compound in the benchmark sets using RDKit (Landrum, 2015). The same fingerprinting method is used to generate fingerprints for each ligand used in the CRANkS grid for a given run. The similarity between the compound and each ligand is calculated in a pairwise fashion using the Tanimoto similarity to calculate the similarity of the two fingerprints (Section 1.3.1).

Morgan fingerprints or extended-connectivity fingerprints (ECFP; Rogers and Hahn, 2010) are designed to capture 2-dimensional chemical features and the connectivity of the molecule. The fingerprinting method is based on the Morgan algorithm (Morgan, 1965) which generates identifiers for each atom that are independent of the numbering of the atoms. For each atom in a molecule, the atom is selected as the central atom. All atoms connected within the radius of bonds are considered. The chemical features of these connected atoms are encoded in the bit string, for example charge, element, whether it is aromatic, whether it is a donor *etc*. Then, the next atom is taken as the central atom and so on until all atoms have been selected. This 2D circular fingerprinting method was chosen as it is similar to ECFP4 which has been

shown to perform well in recent bioactivity coverage analysis (Koutsoukas *et al.*, 2014).

The 3D Pharmacophore fingerprint is generated using the 2D Pharmacophore fingerprint calculator implemented in RDKit but feeding the calculator a 3D distance matrix for the molecule. The pharmacophore fingerprint is calculated by first calculating the pharmacophoric features of the molecule. In this case, the feature library used is the Gobbi feature library in-built in RDKit (Gobbi and Poppinger, 1998). The inter-feature topological distances are then calculated using the 3D distance matrix and a bit id is assigned to each feature-distance combination. These bits then form the fingerprints as shown in Figure 3.10, as adapted from a presentation by Greg Landrum at the RDKit User Group Meeting 2012.

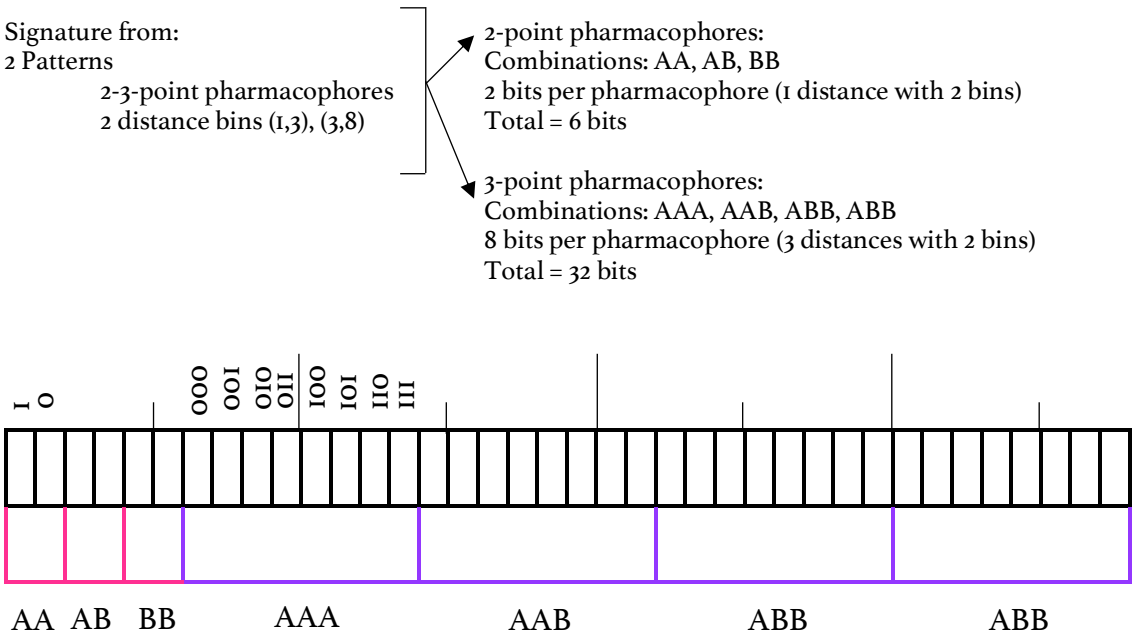



Figure 3.10 Calculation of bits for the 3D Pharmacophore fingerprint calculated using RDKit.

By using the same conformations to calculate the 3D Pharmacophore fingerprints, the 3D Pharmacophore Fingerprint Score can be used as a control for the conformer generation part of the CRANkS algorithm. If both the 3D Pharmacophore Fingerprint Score and the CRANkS algorithm perform badly, this could be due to inaccurate conformer generation. Conversely, if the 3D Pharmacophore Fingerprint Score performs well but the CRANkS algorithm does not this might suggest problems with other parts of the algorithm. The Morgan Fingerprint Score acts as a control for using 3D conformations at all; if the Morgan Fingerprint Score performs well and both the 3D Pharmacophore Fingerprint Score and CRANkS algorithm do not, this could indicate that the addition of 3D information is simply adding noise which outweighs any useful signal. However, this could also indicate problems with the dataset in question. If the 2D fingerprint can easily differentiate the actives and inactives, this could indicate that the inactives are inherently chemically dissimilar to the actives which is not typical of a real drug campaign.

Additional fingerprinting methods can be compared by taking values for testing of the fingerprints on the DUD-E database from the literature. Three 2D fingerprint methods were identified in the literature: SMILES fingerprints (SMIfp) (Schwartz *et al.,* 2013), Molecular Quantum Numbers (MQN) (Nguyen *et al.*, 2009; van Deursen *et al.*, 2010), and $T_{SF}$ (Schwartz *et al.,* 2013). SMIfp is a scalar fingerprint based on the SMILES string that can be used to represent a compound. The occurrence of 34 different symbols in the SMILES string are counted and used to form the fingerprint. MQN is a property-based descriptor that builds the fingerprint based on 42 integer value descriptors of molecular structure which include the count of atom types, bond types, polar groups and topological features. $T_{SF}$ uses the Tanimoto similarity of the

candidate compound with a single ligand using Daylight Fingerprints. The ligand is taken from the protein-ligand crystal structure that DUD-E assigns for each target for docking calculations. MRaise uses molecular alignment with a triangle-descriptor alignment to calculate shape overlap using a Gaussian-type scoring function (von Behren *et al.,* 2016).

### *3.2.5.2 Structure-Based Methods*

By using the DUD-E dataset, calculated metrics for the discrimination of actives and inactives can be compared to other methods which have also been tested on the DUD-E dataset. By searching the literature, a variety of docking and structure-based methods have been found to compare to the CRANkS algorithm.

Four popular commercial docking tools performed a benchmarking study on the whole of the DUD-E database (Chaput *et al.*, 2016). The docking tools tested were GOLD (Jones *et al.,* 1997; Jones *et al.,* 1995), Glide (Friesner *et al.,* 2004; Halgren *et al.*, 2004), Surflex (Jain, 2003) and FlexX (Rarey *et al.*, 1996). All four programs require careful configuration of a number of parameters, which can strongly influence the outcome of the docking protocol. Additionally, AutoDock Vina (Trott and Olsen, 2010), one of the most widely used open-source docking tools was compared to the CRANkS algorithm using the AUC and EF metrics for the DUD-E datasets (Roy *et al.,* 2015).

Three further tools have calculated AUC and EF metrics for the DUD-E datasets (Roy *et al.,* 2015). PoLi uses a generalisation of ligand homology modelling (Roy *et al.,*

2015). It identifies similar binding pockets to the target of interest within a holo template library. Parts of the ligands within these similar binding pockets are then used to perform virtual screening of compounds based on the 2D and 3D similarity of the compounds with the parts of ligands. FINDSITE$^{filt}$ is a homology-based virtual screening tool that uses the 2D similarity of candidate compounds with ligands from protein-ligand structures detected to have a common binding site to the target using a set of evolutionarily related proteins (Zhou and Skolnick, 2013). LIGSIFT uses a Gaussian molecular shape overlay to screen candidate compounds based on 3D shape similarity to a given ligand (Roy *et al.*, 2015).

## 3.3 Results

### *3.3.1 Discrimination of Actives from Inactives*

The results of using CRANkS for the AUC calculated for the benchmark datasets of each target using different numbers of structures are shown in Figure 3.1. The corresponding values for other algorithms as found in the literature for each target are also shown (Roy *et al.,* 2015; Schwartz *et al.,* 2013). The Morgan Fingerprint Score is calculated to have a high AUC for nearly all targets, in particular targets from the DUD-E database (0.63-0.85). This is likely to be because the putative decoys are inherently dissimilar to the actives because of how the DUD-E datasets are constructed. The other 2D fingerprinting methods also do well consistently across the targets: $T_{SF}$ and SMIfp.

Figure 3.11. Table showing calculated AUCs for each target for different methods. For the CRANkS scores and Morgan and 3D Pharmacophore Fingerprint Scores the AUCs are split by the number of input structures, and each calculated AUC is an average over five separate runs of different input structures. The AUCs for the alternative docking methods and fingerprinting methods have been collated from the literature, and the highest AUC of the alternative methods for each target is shown in bold. The highest AUC out of the 3D Pharmacophore Fingerprint Score and each of the CRANkS scores for each set of numbers of input structures is shown in bold. The highest AUC out of all methods for each target is outlined by a black box. White refers to an AUC of 0.5, blue less than 0.5 and pink greater than 0.5. It can be seen from the three targets for which the performance of structure-based methods have been published (ADRB2, BACE1 and GRIA2) that CRANkS performs better than AutoDock Vina, PoLi, LIGSIFT and FINDSITE_filt for these targets.

This result indicates a weakness of the DUD-E datasets. In a drug discovery campaign, it is unlikely that actives could be so easily separated from decoys simply by 2D similarity as they will be developed from the same set of hits as part of hit-to-lead optimisation. Therefore, it should not be taken as an issue that the CRANkS scores do not achieve as high AUCs for the DUD-E datasets. Indeed, AutoDock Vina, a popular open-source docking tool, performs worse than the Morgan Fingerprint Score for all three targets for which there is data. The CRANkS algorithm achieves better results than AutoDock Vina in two out of the three targets for which there are published results to compare which is extremely promising. Additionally, only selecting compounds that are similar to the ligands for which there are already crystal structures will not allow for an adequate exploration of chemical space and would not facilitate scaffold-hopping (see Section 3.3.2) There is no clear increase in performance as the number of structures increases indicating that results can be achieved using a small number of structures.

For the additional datasets that should be more representative of a drug discovery campaign the results are different. For BRD1 and CDK2 datasets the Morgan Fingerprint Score achieves near random AUC values. However, the CRANkS scores show little improvement over these values for all numbers of input structures (for example for CDK2 the AUC for the Morgan Fingerprint Score ranges from 0.51 to 0.54 and the AUC for the CRANkS Interaction Score ranges from 0.53 to 0.57). Unfortunately, there are no additional values to compare these results to from the literature to more clearly decipher whether other tools could achieve higher AUC values for these datasets.

The 3D Pharmacophore Fingerprint Score acts as a conformational control for the CRANkS algorithm as it uses the same conformers. The results of the CRANkS scores are in general similar to the 3D Pharmacophore Fingerprint Score. The CRANkS scores are better than the 3D Pharmacophore Fingerprint Score for ADRB2 (CRANkS Interaction Score average AUC of 0.80 and 3D Pharmacophore Fingerprint Score AUC of 0.69) and HSP90A (e.g. CRANkS Pharmacophore Score AUC of 0.90 and 3D Pharmacophore Score AUC of 0.82 for 50 structures). For targets GRIA2 and DHFR there is a reduction in AUC for the CRANkS scores compared to the 3D Pharmacophore Score indicating the CRANkS algorithm does not achieve as high discrimination as the 3D fingerprinting despite using the same conformations for the candidate compounds.

For targets ADRB2, GRIA2 and BACE1 structure-based methods have also been tested on the sets yielding AUC values that can be compared to those calculated for the CRANkS algorithm. For each of the targets at least one of the CRANkS scores achieved a higher AUC than all of the structure-based methods compared to here: AutoDock Vina, PoLi, LIGSIFT, and FINDSITE$^{filt}$.

The maximum and minimum AUC for each scoring method across the five grids with the same number of input structures is shown in the Appendix in Figure A.3.9. The CRANkS scores exhibit a much larger range of AUC than the fingerprinting scores indicating the sensitivity of the CRANkS scores to the protein-ligand structures used to construct the grids. This is also exhibited by the BEDROC and $EF_{1\%}$ results (Figure A.3.10 and Figure A.3.11).

The results for CRANkS using BEDROC as the metric are more promising (Figure 3.12). The CRANkS algorithm is compared to four of the most widely used commercial docking tools: Surflex, FlexX, GOLD and Glide. For HSP90A the CRANkS algorithm outperforms all four docking tools and for ANDR, BACE1 and PYGM the algorithm outperforms three out of four. For BRD1, the Interaction Score is now calculated to have the best performance, despite a poor AUC. CRANkS scores are only found to have a poorer performance than the docking tools for ADRB2 and GRIA2, and even in these cases the results are not significantly worse. This indicates that the CRANkS algorithm generally performs as well as popular docking tools in terms of early enrichment and discriminating actives from decoys when only using five protein-ligand crystal structures to construct the grid, without requiring a pre-docking protocol.

The results using $EF_{1\%}$ are shown in Figure 3.13. Once again 2D fingerprinting methods achieve the best results for the DUD-E database in the case of the Morgan Fingerprint Score and both $T_{SF}$ and SMIfp found in the literature. However, it should be noted that the CRANkS algorithm consistently outperforms the 3D Pharmacophore Fingerprint Score and, in all cases, outperforms the docking tools.

Figure 3.12. Table showing calculated BEDROC (α = 80.5) values for each target for different methods. For the CRANkS scores and Morgan and 3D Pharmacophore Fingerprint Scores the BEDROC values are split by the number of input structures, and each calculated BEDROC is an average over five separate runs of different input structures. The BEDROCs for the alternative docking methods have been collated from the literature, and the highest AUC of the alternative methods for each target is shown in bold. The highest BEDROC out of the 3D Pharmacophore Fingerprint Score and each of the CRANkS scores for each set of numbers of input structures is shown in bold. The highest BEDROC out of all methods for each target is outlined by a black box.

| | | DUD-E Datasets | | | | | | | Additional Datasets | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | ADRB2 | ANDR | BACE1 | GRIA2 | HSP90A | PYGM | BRD1 | CDK2 | DHFR |
| **5 structures** | Morgan Fingerprint Score | 37.0 | 34.5 | 29.3 | 19.7 | 48.3 | 12.2 | 0.3 | 1.1 | 6.9 |
| | 3D Pharmacophore Fingerprint Score | **24.5** | **17.1** | **27.5** | 12.5 | **45.3** | 5.7 | **0.5** | 1.2 | **10.3** |
| | CRANkS Interaction Score | 15.7 | 6.8 | 12.1 | 9.0 | 22.4 | 0.3 | **0.5** | **2.0** | 0.2 |
| | CRANkS Element Score | 23.8 | 10.7 | 26.9 | 16.6 | 17.0 | **7.0** | 0.2 | 1.8 | 1.2 |
| | CRANkS Pharmacophore Score | 22.4 | 11.5 | 26.1 | **26.3** | 12.6 | 3.4 | 0.2 | 1.7 | 5.3 |
| **10 structures** | Morgan Fingerprint Score | N/A | 36.5 | 46.7 | 24.9 | 36.7 | 12.5 | 0.3 | 1.3 | 9.6 |
| | 3D Pharmacophore Fingerprint Score | N/A | **21.0** | **41.7** | 10.8 | **32.0** | **6.5** | 0.5 | 1.2 | **9.8** |
| | CRANkS Interaction Score | N/A | 6.8 | 3.6 | 5.6 | 30.0 | 0.0 | **0.6** | 0.6 | 0.8 |
| | CRANkS Element Score | N/A | 5.2 | 41.3 | 14.2 | 20.4 | **6.5** | 0.1 | **1.8** | 0.4 |
| | CRANkS Pharmacophore Score | N/A | 8.9 | 38.9 | **15.0** | 18.9 | 4.4 | 0.1 | 1.5 | 2.4 |
| **25 structures** | Morgan Fingerprint Score | N/A | N/A | 53.6 | 31.3 | 45.8 | 11.3 | 0.2 | **2.9** | 10.2 |
| | 3D Pharmacophore Fingerprint Score | N/A | N/A | 47.0 | **12.5** | 36.9 | 6.3 | 0.5 | 0.9 | **6.5** |
| | CRANkS Interaction Score | N/A | N/A | 15.9 | 10.0 | 23.2 | 0.0 | 0.5 | 1.1 | 1.3 |
| | CRANkS Element Score | N/A | N/A | **47.8** | 10.5 | 31.5 | **7.3** | 0.3 | 1.7 | 0.8 |
| | CRANkS Pharmacophore Score | N/A | N/A | 47.7 | 10.8 | **49.7** | **7.3** | 0.2 | **2.9** | 2.6 |
| **50 structures** | Morgan Fingerprint Score | N/A | N/A | N/A | N/A | 57.3 | 12.4 | N/A | 2.6 | 6.1 |
| | 3D Pharmacophore Fingerprint Score | N/A | N/A | N/A | N/A | 39.2 | 6.5 | N/A | 1.1 | **7.4** |
| | CRANkS Interaction Score | N/A | N/A | N/A | N/A | 36.8 | 0.0 | N/A | 1.1 | 2.6 |
| | CRANkS Element Score | N/A | N/A | N/A | N/A | 37.5 | 7.0 | N/A | 1.3 | 0.0 |
| | CRANkS Pharmacophore Score | N/A | N/A | N/A | N/A | **55.6** | **7.6** | N/A | **3.0** | 5.7 |
| | Theortical Maximum | 65.9 | 54.3 | 65.0 | 76.0 | 62.8 | 52.3 | 2.5 | 59.8 | 15.1 |
| **Structure-Based Methods** | AutoDock Vina | 3.0 | N/A | 4.6 | 2.5 | N/A | N/A | N/A | N/A | N/A |
| | PoLi | 2.6 | N/A | 4.2 | 7.0 | N/A | N/A | N/A | N/A | N/A |
| | LIGSIFT | 1.7 | N/A | 3.5 | 5.7 | N/A | N/A | N/A | N/A | N/A |
| | FINDSITE_filt | 0.4 | N/A | 11.3 | 1.3 | N/A | N/A | N/A | N/A | N/A |
| **Fingerprinting Methods** | MRaise | 19.1 | 19.4 | 2.5 | 46.6 | 32.1 | 1.3 | N/A | N/A | N/A |
| | Tsf | **34.0** | **20.0** | **39.9** | 45.5 | **46.0** | **20.4** | N/A | N/A | N/A |
| | SMIfp | 22.0 | 6.3 | 12.7 | 12.0 | 29.2 | 19.1 | N/A | N/A | N/A |
| | MQN | 12.9 | 17.4 | 8.1 | 9.5 | 31.4 | 10.2 | N/A | N/A | N/A |

EF: 0.0 — Theoretical Maximum

Figure 3.13. Table showing calculated $EF_{1\%}$ for each target for different methods. For the CRANkS scores and Morgan and 3D Pharmacophore Fingerprint Scores the $EF_{1\%}$ values are split by the number of input structures, and each calculated $EF_{1\%}$ is an average over five separate runs of different input structures. The $EF_{1\%}$ values for the alternative docking methods and fingerprinting methods have been collated from the literature, and the highest $EF_{1\%}$ of the alternative methods for each target is shown in bold. The highest $EF_{1\%}$ out of the 3D Pharmacophore Fingerprint Score and each of the CRANkS scores for each set of numbers of input structures is shown in bold. The highest $EF_{1\%}$ out of all methods for each target is outlined by a black box.

Using the AUC, BEDROC and $EF_{1\%}$ metrics to calculate the ability of active versus inactive discrimination I have found that although the performance is target-specific, the CRANkS algorithm can differentiate as well as if not better than popular docking and screening methods. Notably the algorithm performed well in initial enrichment and was found to outperform 3 out of 4 commercial docking tools using the BEDROC metric for most targets.

### *3.3.1.1 Individual Targets*

#### *3.3.1.1.1 Beta-2 Adrenergic Receptor - ADRB2*

ADRB2 (beta-2 adrenergic receptor) is a cell membrane spanning G-protein coupled receptor. A conformational change upon binding induces a signal for smooth muscle relaxation. Agonists have been developed against this target for asthma, for example salbutamol, or for pulmonary disorders. There were seven crystal structures available (Table 3.1) and the ligands of the collated crystal structures are shown in Figure 3.14. As the conformational change occurs upon binding the proteins are all in the same conformation and the ligands are well-aligned.

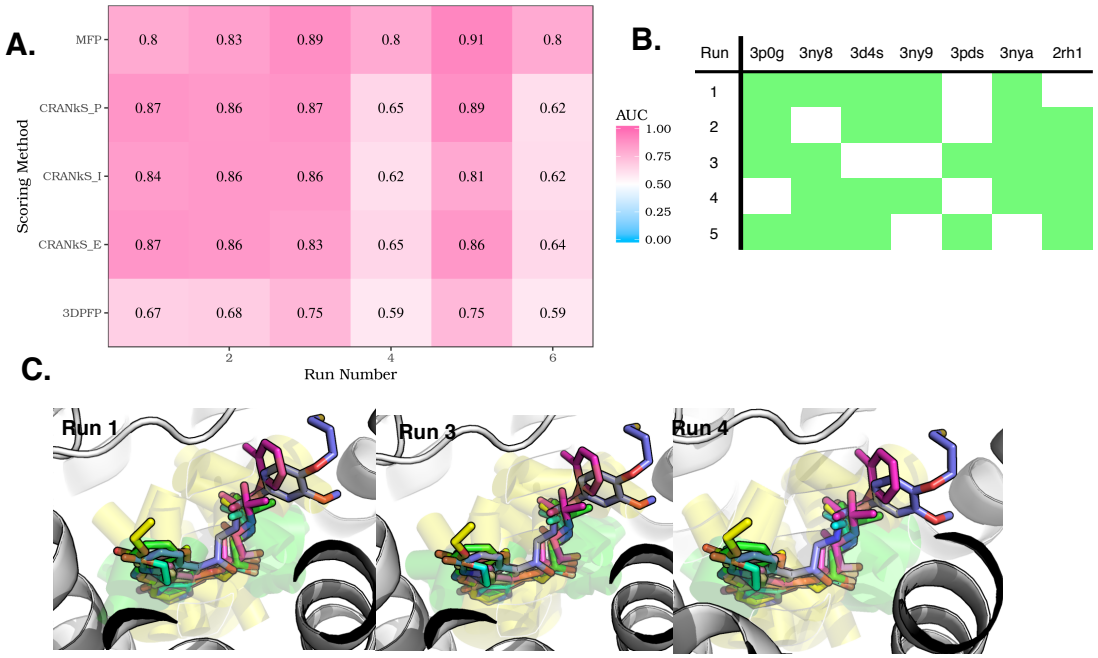



Figure 3.14. The AUC for runs of 5 protein-ligand structures used to form the CRANkS grids for ADRB2 are shown in A. Run 6 corresponds to the conformations from run 4 rescored with the grid from run 3. The protein-ligand complexes used to form the grids in each run are shown in B (green if used and white if not). The ligands within the binding pockets are well-aligned as shown in C. Notably there is little difference between Run 1, 3 and 4 in terms of the interaction grids, despite the differences in the protein-ligand structures used.

The AUC results for the runs of 5 randomly selected structures are shown in Figure 3.14. Notably one run gets substantially worse results - run 4, for the CRANkS algorithm scores and also the 3D Pharmacophore Fingerprint score (Figure 3.14 A). In comparison the Morgan Fingerprint Score does not score significantly worse. To investigate the cause of this Figures 3.14 C shows the structures in each run. Interestingly, run 4 does not include the crystal structure *3p0g* whereas all other runs do. The interaction grids with the ligands from the crystal structures are shown in Figure 3.14 C. The interaction grid for run 4 is not significantly different to the other runs. The conformations in run 4 were rescored using the grids calculated in run 3 to investigate whether it is the grid or conformations causing the reduction in performance. The results from the rescored run (shown as run 6) were no better than for the original run 4, indicating that the issue lies with the conformer generation.

The ligand in the PDB structure 3p0g has an additional aromatic ring that points away from the other ligands. It is likely then that this ring is used to place conformers for the actives and decoys when tested, and the removal of this anchor is what causes the CRANkS algorithm to not be able to discriminate between actives and decoys sufficiently well for this run. This also explains why the 3D Pharmacophore Fingerprint Score has a significantly worse AUC for this run as it uses the same conformers as the CRANkS algorithm. This highlights how sensitive the 3D MMP conformer generation method is to the crystal structures used as input. It would be prudent to explore how to improve conformer generation and whether to use another method such as docking or constrained docking.

#### *3.3.1.1.2 GRIA2*

GRIA2 (glutamate ionotropic receptor AMPA type subunit 2) is a non-selective cation channel that allows for the passage of sodium and potassium ions across the cell membrane. Post-transcriptional modification also allows for calcium ions to pass. Upon ligand binding, the structure contracts inducing domain closure of the transmembrane gates. The extent of the receptor activation depends on the degree of the domain closure. This is a potential target for therapies for the central nervous system disorders including epilepsy and Alzheimer's disease. From the DUD-E database, 80 crystal structures of protein-ligand complexes were collated from the PDB with a matching UniProt code to the target. However, after visual inspection only 30 structures were used in this analysis. Structures were removed that contain the same compound in the same orientation or structures with the ligand bound in an allosteric site. Five randomly chosen sets of 5, 10 and 25 structures were used to

construct CRANkS grids for this analysis. 12003 molecules from the DUD-E dataset were scored of which 158 were actives and 11845 were decoys.

The calculated AUC, BEDROC and $EF_{1\%}$ are shown for each number of randomly selected input structures in Figure 3.15. The mean value across the 5 runs for each size of input structures is shown with the standard deviation indicated by an error bar. For grids constructed using five structures the Morgan Fingerprint Score and 3D Pharmacophore Fingerprint Score do not significantly outperform the CRANkS algorithm for any of the three measures (Figure 3.15). As the number of input structures increases the AUC, BEDROC and $EF_{1\%}$ values calculated for the Morgan Fingerprint Score increases: adding more structures is beneficial, and this is consistent across all measures. The AUC of the 3D Pharmacophore Fingerprint Score also increases which suggests that adding more input structures does not just increase noise in the calculation but adds value to the conformer generation step. However, the measured BEDROC and $EF_{1\%}$ for the 3D Pharmacophore Fingerprint Score remains very similar across the numbers of input structures indicating that although the separation of actives from decoys improves, early recognition stays similar. However, notably for both BEDROC and $EF_{1\%}$ of the 3D Pharmacophore Fingerprint Score the standard deviation decreases with increasing number of structures. This could indicate that adding more structures adds stability to the conformer generation step rather than simply adding noise.

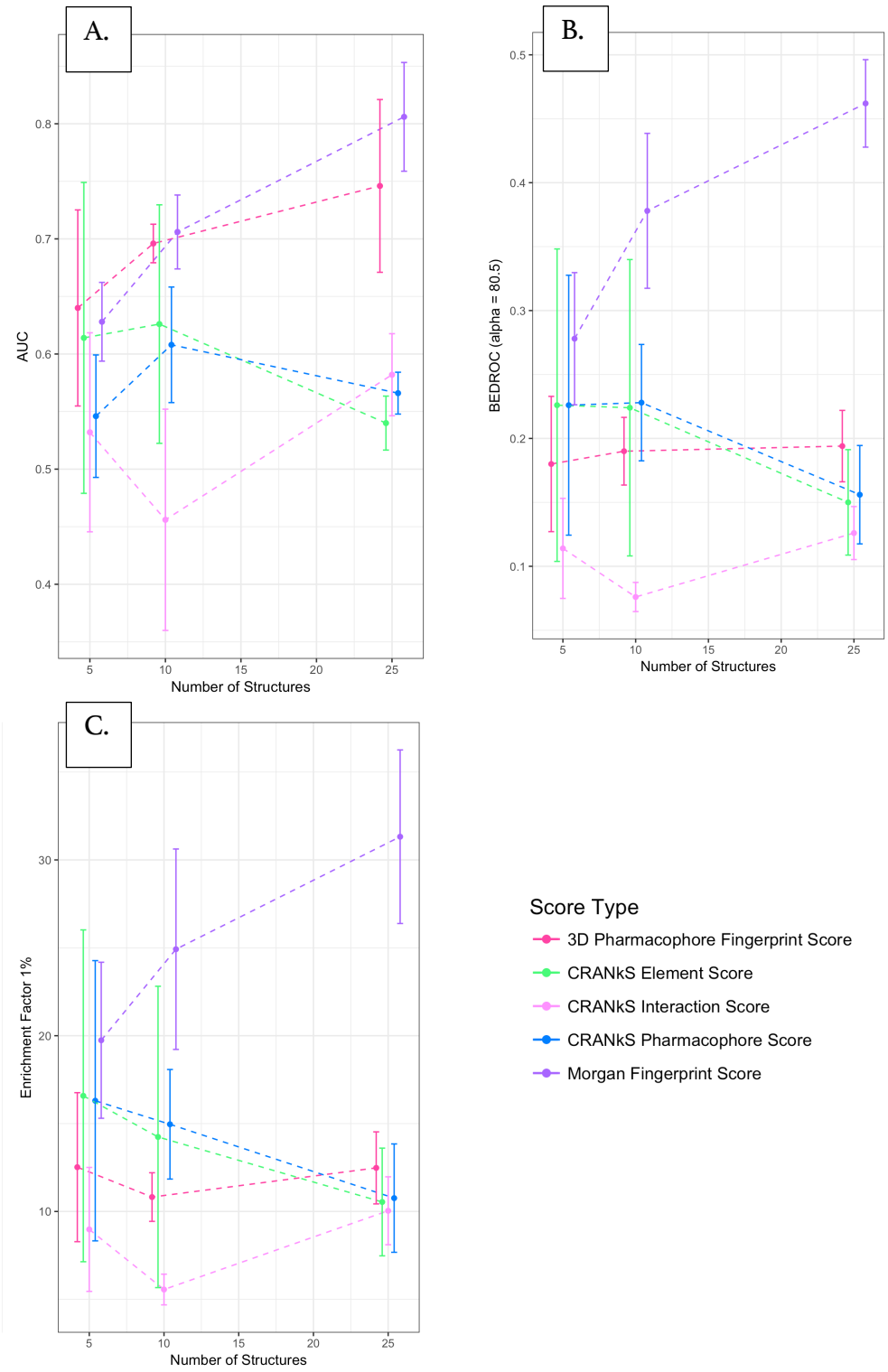



Figure 3.15. Results of the AUC (A), BEDROC (B) and $EF_{1\%}$ (C) as a function of the number of input structures for GRIA2. There is little variation in the mean of each metric calculated for the CRANkS scores as the number of structures increases. However, the standard deviation decreases with an increasing number of structures.

In the case of the CRANkS scores, across all three metrics, there is a reduction in value as the number of input structures increases. However, the calculated values are all within one standard deviation of each other, *i.e.* the calculation for 25 structures is within one standard deviation of the calculation for 10 structures, which in turn is within one standard deviation for 5 structures. Although adding more input structures does not appear to aid the separation of actives from decoys, it is also not conclusive whether the separation is worsened by an increase in noise. There is a very large deviation for all measures on the CRANkS scores when 5 input structures are used. This indicates the algorithm's sensitivity to the protein-ligand complexes used to construct the grid. Across all three measures there is a clear decrease in the standard deviation for all of the CRANkS scores upon increasing the number of structures used to construct the CRANkS grids. This indicates that the score stabilises with more input structures but does not necessarily improve performance.

Reported AUCs, BEDROC and $EF_{1\%}$ values from other algorithms for this dataset collated from the literature are shown in Figures 3.11, 3.12 and 3.13. For this target, the AUC of AutoDock Vina outperforms both the CRANkS algorithm and fingerprinting tools when only 5 input structures were used (AUCs for AutoDock Vina = 0.61, CRANkS Interaction Score = 0.53, CRANkS Element Score = 0.62, CRANkS Pharmacophore Score = 0.55, Morgan Fingerprint Score = 0.63 and 3D Pharmacophore Score = 0.64). However, the fingerprinting tools outperform AutoDock Vina when the number of structures increases to 10 and 25, whereas the CRANkS algorithm still underperforms. However, in the case of $EF_{1\%}$ AutoDock Vina is outperformed by both the CRANkS algorithm and the fingerprinting methods. In fact, at least two of the CRANkS scores outperform all additional

methods as reported in the literature (with the exception of MRaise and $T_{SF}$ for the $EF_{1\%}$ calculated for both 5 and 10 input structures). This indicates the CRANkS algorithm has early enrichment comparative to other structure-based methods. In the case of BEDROC, both the Pharmacophore and Element scores have higher results than FlexX and Surflex: two popular commercial docking tools. However, both Gold and Glide docking tools have higher calculated BEDROC than the CRANkS algorithm for all numbers of input structures.

It is clear that for this target the CRANkS algorithm underperforms in the separation of actives from decoys when using AUC as the measurement and compared to methods both from the literature and the fingerprinting scores calculated as part of this work. However, the $EF_{1\%}$ of separating from actives and decoys was found to be similar to that of other widely-used methods indicating the CRANkS algorithm achieved adequate early recognition for this target.

The fact the 3D Pharmacophore Fingerprint Score achieves better results than for the CRANkS algorithm indicates that the issue for this target (GRIA2) is not due to the conformer generation part of the algorithm. It is therefore likely to be due to the structures of the protein or the alignment of the ligands in the grids. This target changes conformation on ligand binding and the degree of conformational change affects the function of the target. Thus, different protein-ligand structures in the set have different binding site conformations. This is illustrated in Figure 3.16 which shows the binding sites for each of the sets of 5 input structures. One of the loops has clearly different distances from the ligand for each structure and this varies across the runs. This would have a great effect on the interaction grid as interactions with

different distances would not overlap. It could also have an effect on the pharmacophore and element grids, as there are effectively multiple different binding pockets, which could cause ligands to be misaligned.

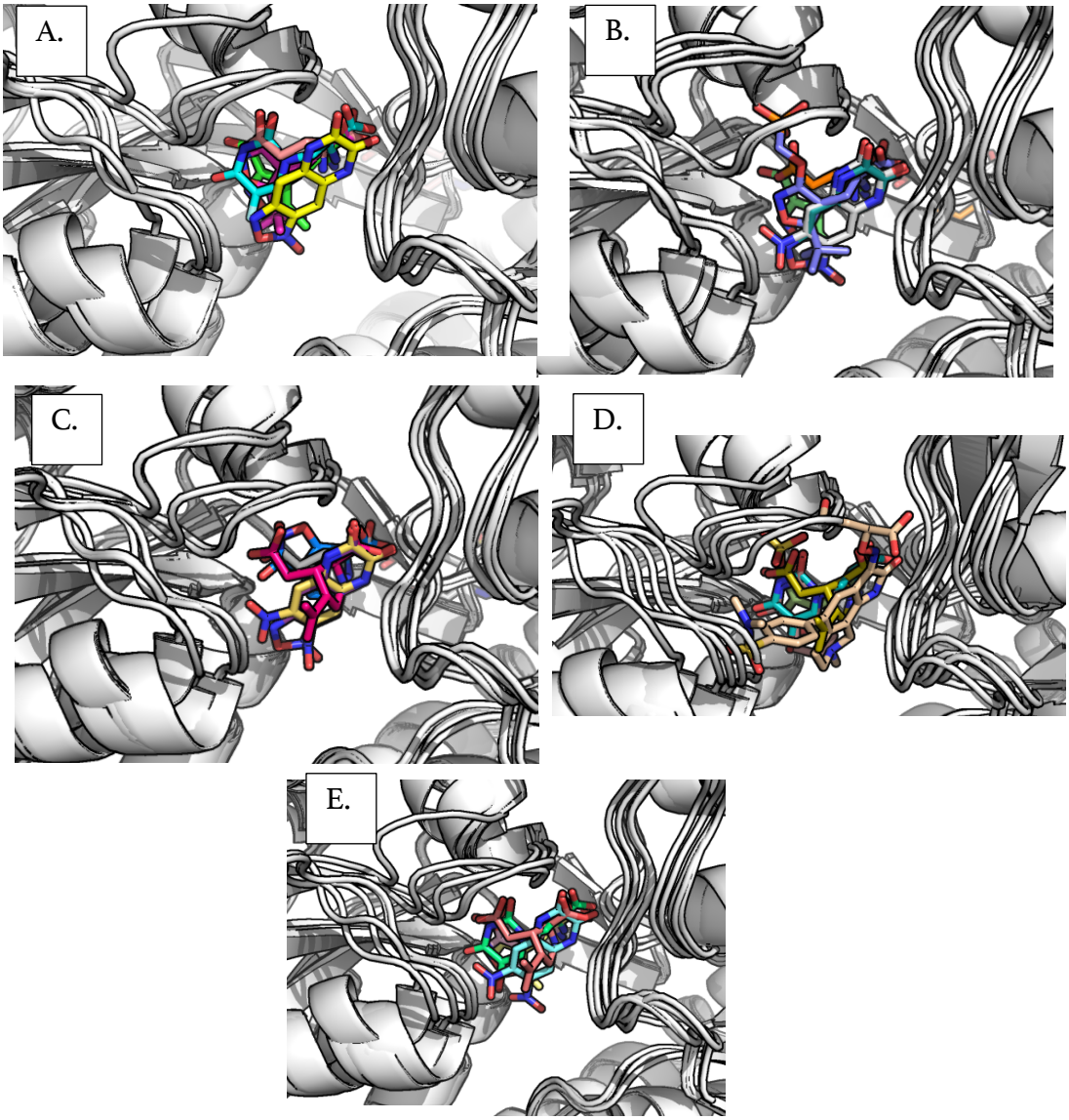



Figure 3.16. The protein-ligand structures used as input for each runs 1 (A), 2 (B), 3(C), 4(D), and 5 (E) for sets of five structures for GRIA2. There are clearly differing conformations and big movements in one of the loops which could be causing the weak performance of the CRANkS algorithm. The PDB structures used to construct each grid are listed in Appendix B, Table B.3.44 to Table B.3.48.

### *3.3.1.2 The Effect of Protein-Ligand Alignment*

The results for the calculated AUC, BEDROC and $EF_{1\%}$ for the datasets are target-dependent. The standard deviations of these values also indicate the sensitivity of the CRANkS algorithm on the input structures (this is discussed in the case of ADRB2 and GRIA2 in Section 3.3.1.1), as well as the maximum and minimum achieved for the metric for each target over the five grids (Appendix A, Figures A.3.9 to A.3.11). By inspection, the targets with datasets where proteins adopt the same conformation and ligands are well-aligned within one region of the binding site are likely to perform well when using the CRANkS algorithm. Examples of this include HSP90A and ADRB2 (Figure 3.17 A and B). On the other hand, targets where structures show multiple conformations or ligands that span a large area of the binding site yield a poor performance from the CRANkS algorithm. Example of this include GRIA2 and PYGM (Figure 3.17 C and D). I investigated the dependency of this by exploring the alignment of the ligands themselves, and also the alignments of the protein structures.

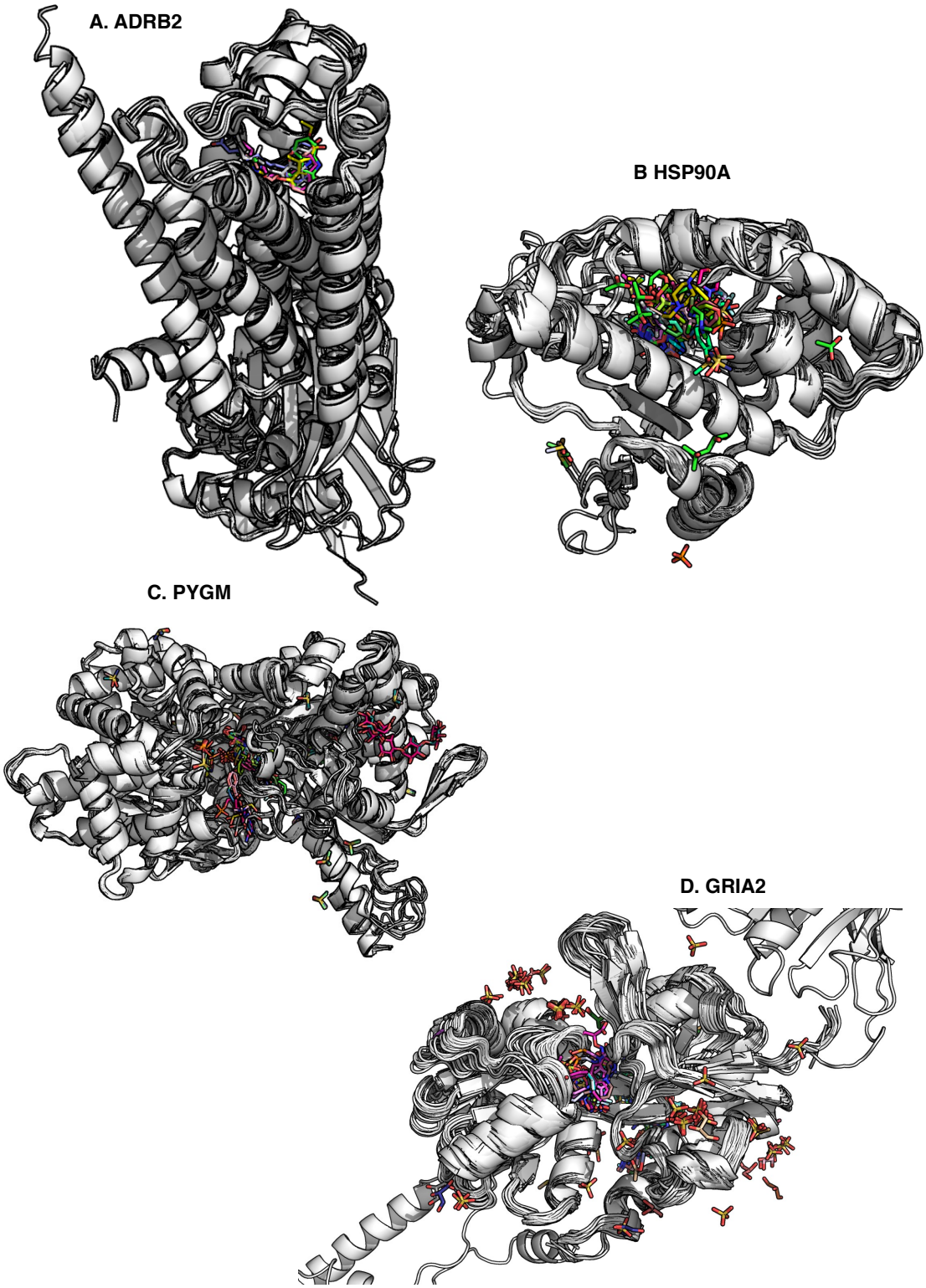
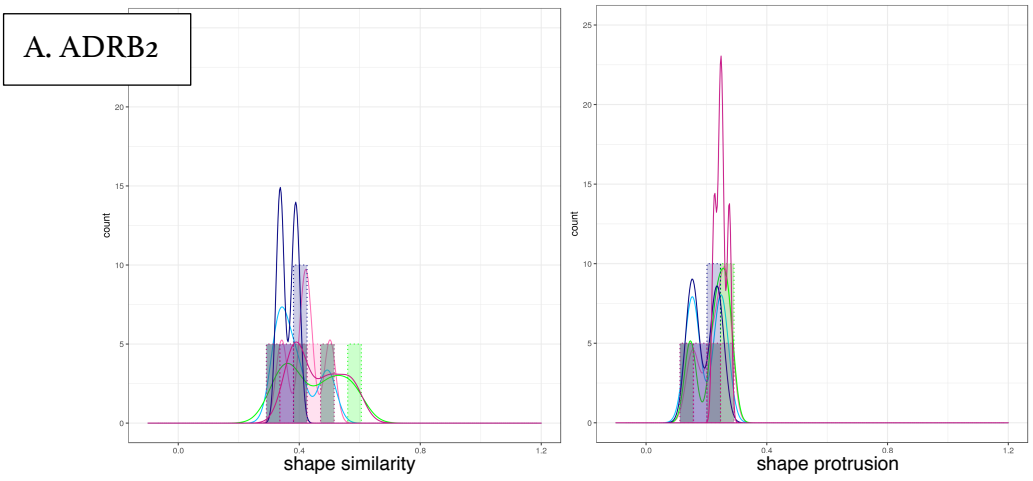



Figure 3.17 Protein-Ligand Structures for four targets used in this analysis: A: ADRB2, B: HSP90A, C: PYGM and D: GRIA2. Notably the ligands are well- aligned for ADRB2 and the protein conformations are very similar for both ADRB2 and HSP90A. This is coupled with high performance of the algorithm. In contrast the ligands are spread out across a much larger site for PYGM. For GRIA2 the proteins are in multiple conformations and are not well aligned in the binding site. This is coupled with poor performance of the algorithm for these targets.

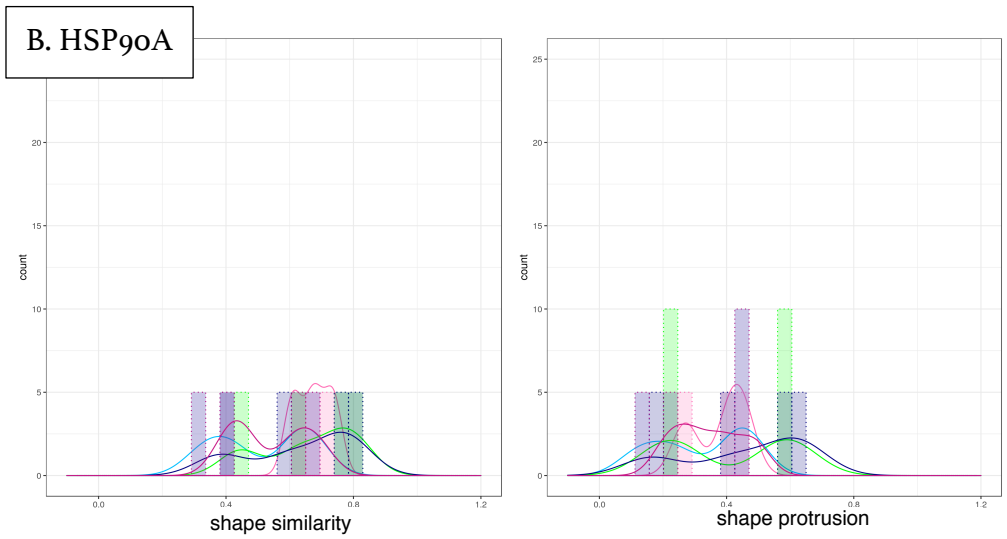
The alignment of the ligands was investigated using shape protrusion and shape similarity metrics calculated using RDKit. The values of these metrics vary from zero to one, where a value of one would correspond to perfect overlap of both shapes. The shape protrusion uses the volume mismatch between two molecules to calculate the similarity between two 3D conformations molecules, whereas the shape similarity uses the volume overlay. To investigate the effect of ligand alignment on the performance of the algorithm, the pairwise similarities for all the ligands used in a grid were calculated. The distributions of these for the five runs of 25 protein-ligand structures for HSP90A, ADRB2, GRIA2 and PYGM are shown in Figure 3.18. The AUCs for the respective runs are also shown.

The distribution for ADRB2 (Figure 3.18 A) shows the ligands are well-aligned with a very tight distribution. This corresponds to consistently high AUCs. The AUC for run 4 is lower than the other the runs. There does not appear to be a clear difference in the shape protrusion distributions. However, for the shape similarity distributions, run 4 shows a distribution skewed towards a lower similarity, which could correspond to the absence of *3p0g* described in Section 3.3.1.1. The distributions for PYGM (Figure 3.18 C) show that some of the molecules did not overlap at all, with a score of 1.0 for the shape protrusion and shape similarity. These runs are runs 4 and 5 which perform much worse than the other runs. However, HSP90A shows a broader distribution of similarities whilst still maintaining AUC performance (Figure 3.18 B). In fact, the distributions for GRIA2 (Figure 3.18 D) and HSP90A (Figure 3.18 B) are not dissimilar indicating that in the case of GRIA2 ligand overlap is unlikely to be the cause of low performance.




| | Run 1 | Run 2 | Run 3 | Run 4 | Run 5 |
|---|---|---|---|---|---|
| **Morgan Fingerprint Score** | 0.80 | 0.83 | 0.89 | 0.80 | 0.91 |
| **CRANkS Pharmacophore Score** | 0.87 | 0.86 | 0.87 | 0.65 | 0.89 |
| **CRANkS Interaction Score** | 0.84 | 0.86 | 0.86 | 0.62 | 0.81 |
| **CRANkS Element Score** | 0.87 | 0.86 | 0.83 | 0.65 | 0.86 |
| **3D Pharmacophore Fingerprint Score** | 0.67 | 0.68 | 0.75 | 0.59 | 0.75 |

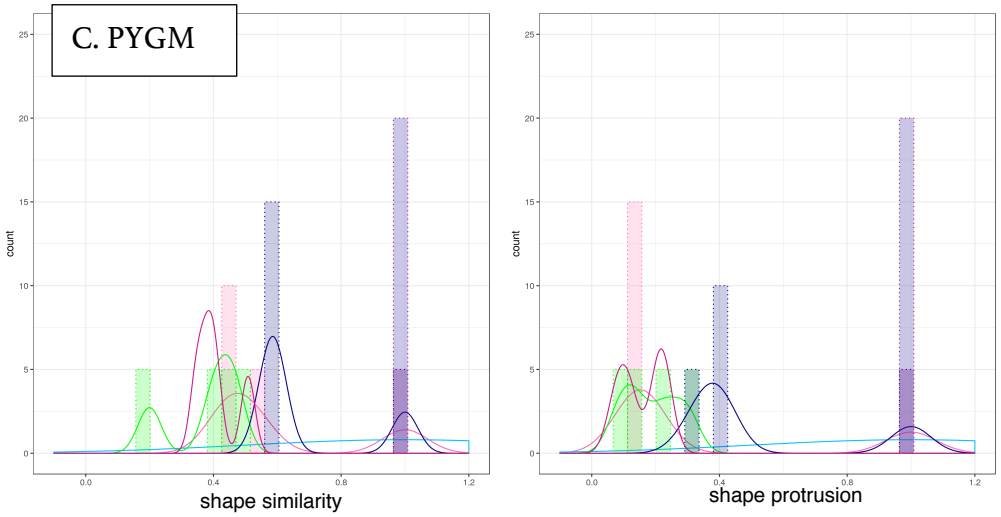
AUC 1.0 0.5 0.0



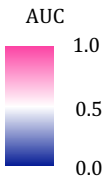

| | Run 1 | Run 2 | Run 3 | Run 4 | Run 5 |
|---|---|---|---|---|---|
| **Morgan Fingerprint Score** | 0.98 | 0.90 | 0.61 | 0.88 | 0.93 |
| **CRANkS Pharmacophore Score** | 0.60 | 0.69 | 0.65 | 0.68 | 0.84 |
| **CRANkS Interaction Score** | 0.88 | 0.69 | 0.69 | 0.92 | 0.56 |
| **CRANkS Element Score** | 0.76 | 0.77 | 0.60 | 0.72 | 0.87 |
| **3D Pharmacophore Fingerprint Score** | 0.87 | 0.83 | 0.81 | 0.83 | 0.87 |

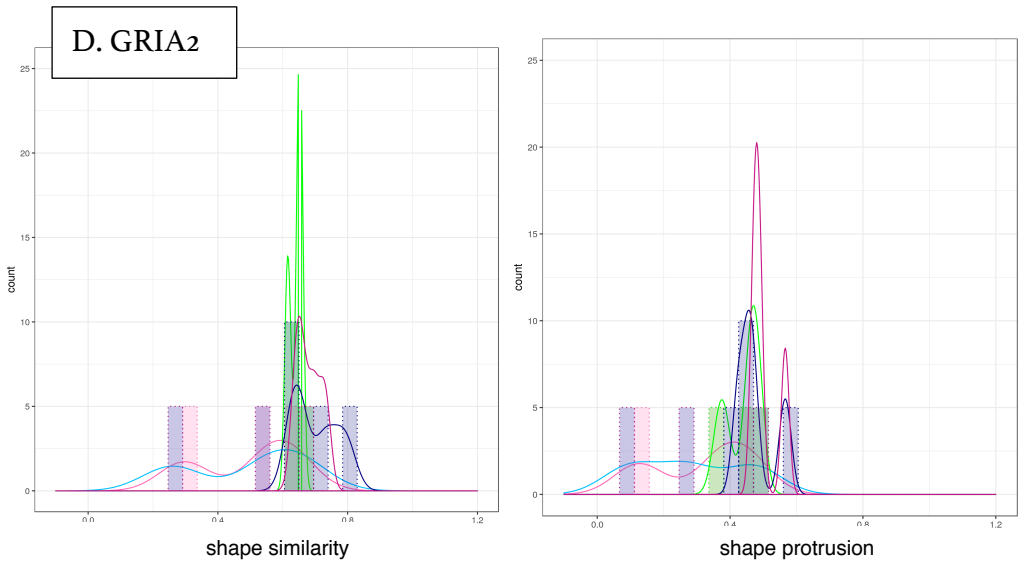
AUC 1.0 0.5 0.0

Figure 3.18. *Continued on next page.*




| | Run 1 | Run 2 | Run 3 | Run 4 | Run 5 |
|---|---|---|---|---|---|
| **Morgan Fingerprint Score** | 0.59 | 0.68 | 0.62 | 0.61 | 0.64 |
| **CRANkS Pharmacophore Score** | 0.59 | 0.61 | 0.48 | 0.53 | 0.52 |
| **CRANkS Interaction Score** | 0.55 | 0.64 | 0.52 | 0.40 | 0.55 |
| **CRANkS Element Score** | 0.71 | 0.76 | 0.60 | 0.59 | 0.41 |
| **3D Pharmacophore Fingerprint Score** | 0.64 | 0.66 | 0.55 | 0.77 | 0.58 |




| | Run 1 | Run 2 | Run 3 | Run 4 | Run 5 |
|---|---|---|---|---|---|
| **Morgan Fingerprint Score** | 0.96 | 0.84 | 0.81 | 0.81 | 0.83 |
| **CRANkS Pharmacophore Score** | 0.59 | 0.63 | 0.52 | 0.43 | 0.45 |
| **CRANkS Interaction Score** | 0.58 | 0.69 | 0.48 | 0.47 | 0.43 |
| **CRANkS Element Score** | 0.68 | 0.66 | 0.55 | 0.36 | 0.43 |
| **3D Pharmacophore Fingerprint Score** | 0.59 | 0.56 | 0.56 | 0.58 | 0.60 |

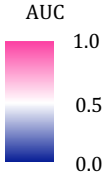



Figure 3.18. Distributions and density functions for the shape similarity and shape protrusion for the ligands of 25 protein-ligand complexes used in 5 separate runs of CRANkS grids. Each colour corresponds to a separate run. The corresponding AUC values for each run are also shown. This is shown for ADRB2 (A), HSP90A (B), PYGM (C) and GRIA2 (D). Run 1 is dark blue, run 2 is pink, run 3 is dark pink, run 4 is green and run 5 is light blue.

Therefore, the conformations of the protein used to construct the grids for HSP90A and GRIA2 were investigated. The binding site was visually determined and the pairwise RMSD for all atoms in the binding site was calculated for the protein structures used in the CRANkS grids for these targets. The distributions of the pairwise RMSD for all atoms, for CRANkS grids containing 25 structures is shown in Figure 3.19. The distributions for HSP90A are all similar with two peaks at approximately 1 Å and 2 Å (Figure 3.19 A. C and E). These are tight distributions and the proteins are consequently in similar conformations. However, the distributions for GRIA2 are much wider, with RMSD values as large as 20 Å (Figure 3.19 B, D and F). The only run which does not have calculated RMSDs this large is run 2 (Figure 3.19 D) and in Figure 3.18 this run is shown to have the best performance as measured by AUC. This suggests that the conformation of the proteins could be causing the poor performance of the CRANkS algorithm for GRIA2. Further work would be required to confirm this hypothesis. The protein-ligand structures should be clustered by the RMSD of the binding site. The CRANkS algorithm should then be run using the clusters as input for the grid to see if there is an increase in performance when proteins of the same conformation are used.

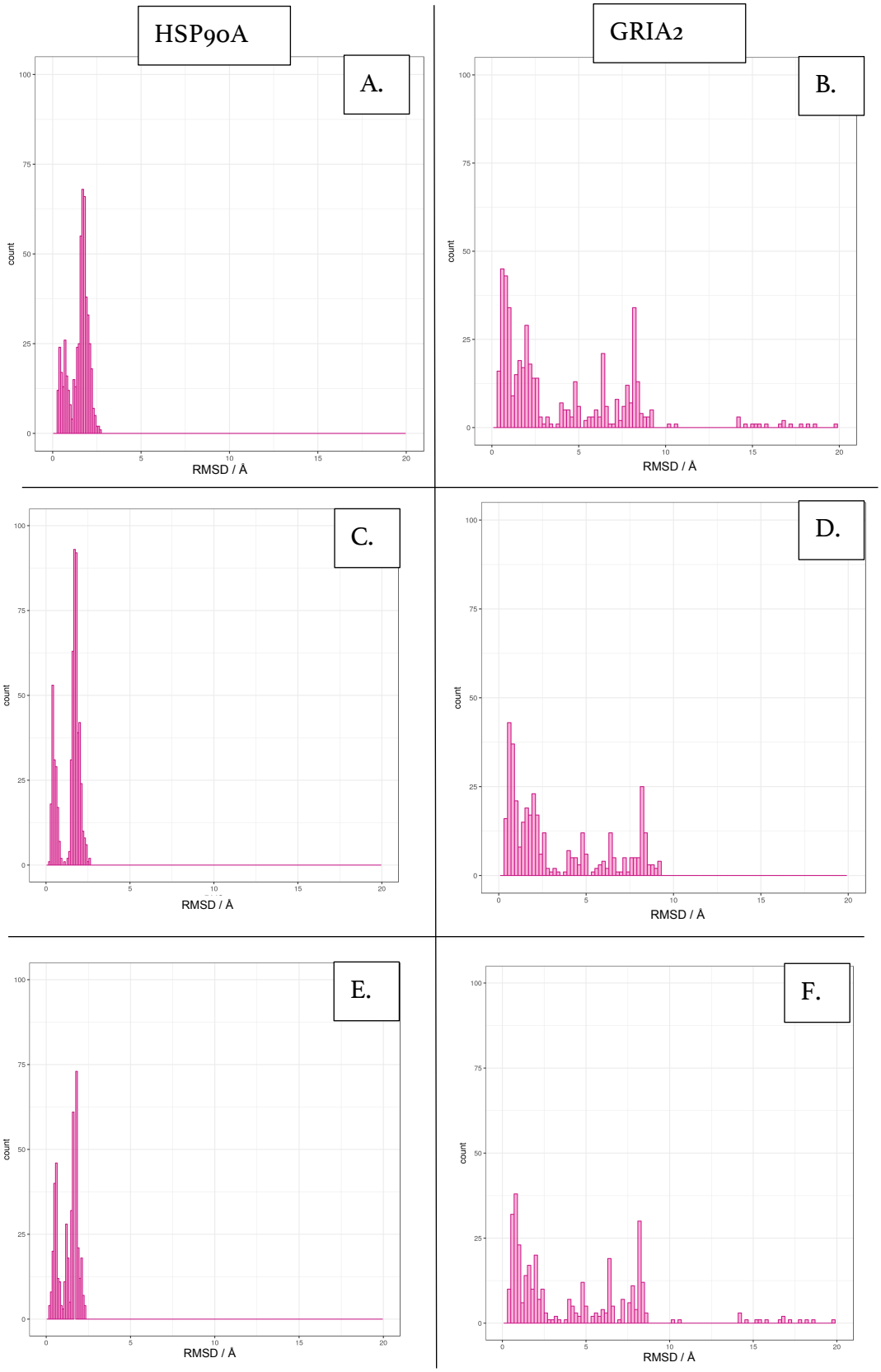



Figure 3.19 Distributions of the pairwise RMSD for the binding site defined by visual inspection for protein structures used in the CRANkS algorithms for HSP90A and GRIA2. Three runs of the 25 structure grids are shown: run 1 (A – HSP90A, B – GRIA2), run 2 (C– HSP90A, D –GRIA2) and run 3 (E- HSP90A and F- GRIA2).

## *3.3.2 Scaffold-Hopping Potential*

### *3.3.2.1 Scaffold-Diversity Metrics*

The mean number of molecules required to recover 50% of the unique scaffolds when ranked by each score, for each number of input structures, for each target, is shown in Figure 3.20. In this case both the active and inactive scaffolds were used. The fewer the number of molecules required to find half of the unique scaffolds, the more unique scaffolds are prioritised, which suggests a higher potential for scaffold hopping due to a greater scaffold diversity. The fewest molecules required for each number of structures for each target is shown in bold. The fewest molecules required for each target across all number of input structures is outlined in black. Notably for GRIA2, HSP90A, CDK2 and DHFR one of the CRANkS scores requires the fewest molecules to reach 50% of the unique scaffolds for all number of input structures. All three of the CRANkS scores require fewer molecules to reach 50% of the scaffolds than the two fingerprinting scores for these targets. In the case of PYGM and BRD1, the CRANkS Interaction Score requires the fewest molecules for 5 protein-ligand structures and also for 10 protein-ligand structures in the case of BRD1. However, as the number of input structures increases the ranking by the Morgan Fingerprint Score requires significantly fewer molecules to gain 50% of the scaffolds. In the case of BRD1, whilst for 25 input structures the Morgan Fingerprint Score on average requires significantly fewer molecules to achieve 50% of the scaffolds, this also corresponds to an AUC that is notably lower than for the other scoring methods. In contrast for PYGM, the AUC for the Morgan Fingerprint Score is extremely high despite the higher scaffold diversity acquired.

| | | DUD-E Datasets | | | | | | Additional Datasets | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | ADRB2 | ANDR | BACE1 | GRIA2 | HSP90A | PYGM | BRD1 | CDK2 | DHFR |
| **5 structures** | Morgan Fingerprint Score | 4364 | 4311 | 4686 | 2830 | 1244 | 843 | 360 | 993 | 634 |
| | 3D Pharmacophore Fingerprint Score | 5743 | 4531 | 4336 | 3220 | 1434 | 829 | 464 | 1226 | 576 |
| | CRANkS Interaction Score | 3191 | **3516** | **3601** | **2547** | 1122 | **804** | **257** | **558** | **206** |
| | CRANkS Element Score | 3873 | 4278 | 3935 | 2845 | 1054 | 949 | 378 | 627 | 280 |
| | CRANkS Pharmacophore Score | **3168** | 3528 | 3612 | 2602 | **998** | 941 | 363 | 645 | 252 |
| **10 structures** | Morgan Fingerprint Score | N/A | 4157 | 4689 | 3013 | 1228 | **782** | 319 | 870 | 662 |
| | 3D Pharmacophore Fingerprint Score | N/A | 4538 | 4864 | 3582 | 1647 | 844 | 402 | 1138 | 570 |
| | CRANkS Interaction Score | N/A | 3580 | 3580 | 2711 | 1173 | 842 | **297** | **608** | 246 |
| | CRANkS Element Score | N/A | 4082 | 4086 | 2711 | 1082 | 913 | 461 | 700 | 256 |
| | CRANkS Pharmacophore Score | N/A | **3437** | **3537** | **2462** | **982** | 958 | 417 | 757 | **239** |
| **25 structures** | Morgan Fingerprint Score | N/A | N/A | 4689 | 2370 | 1244 | **724** | **210** | 1028 | 676 |
| | 3D Pharmacophore Fingerprint Score | N/A | N/A | 5083 | 2633 | 1574 | 865 | 324 | 1062 | 590 |
| | CRANkS Interaction Score | N/A | N/A | **3280** | **2218** | 1069 | 863 | 287 | 600 | **341** |
| | CRANkS Element Score | N/A | N/A | 4078 | 2516 | 1268 | 896 | 427 | 696 | 371 |
| | CRANkS Pharmacophore Score | N/A | N/A | 3754 | 2418 | **971** | 954 | 487 | **514** | 354 |
| **50 structures** | Morgan Fingerprint Score | N/A | N/A | N/A | N/A | 1257 | **743** | N/A | 990 | 676 |
| | 3D Pharmacophore Fingerprint Score | N/A | N/A | N/A | N/A | 1564 | 848 | N/A | 1003 | 600 |
| | CRANkS Interaction Score | N/A | N/A | N/A | N/A | 1092 | 809 | N/A | **520** | **378** |
| | CRANkS Element Score | N/A | N/A | N/A | N/A | 1210 | 834 | N/A | 789 | 466 |
| | CRANkS Pharmacophore Score | N/A | N/A | N/A | N/A | **971** | 1031 | N/A | 565 | 448 |
| | Minimum Number of Molecules | 180.5 | 302 | 219 | 258 | 153 | 138 | 39 | 37 | 9 |
| | Total Number of Scaffolds | 361 | 603 | 439 | 516 | 305 | 275 | 77 | 74 | 17 |
| | Total Number of Molecules | 15231 | 14619 | 18383 | 12003 | 4938 | 4027 | 1353 | 3465 | 756 |




Figure 3.20. Table to showing the number of molecules required for 50% of the scaffolds. This is calculated for each target for different methods. For the CRANkS scores and Morgan and 3D Pharmacophore Fingerprint Scores the values are split by the number of input structures, and each calculated value is an average over five separate runs of different input structures. The lowest number of molecules required for each number of input structures is shown in bold, and the lowest number of molecules required across all numbers of input structures is shown with a black outline.

To look more closely at the initial scaffold diversity of molecules ranked by each scoring method, the average number of scaffolds in the top 100 molecules for each scoring method, for each number of input structures for each target is shown in Figure 3.21. The results for this calculation are more varied within targets and across targets. For HSP90A, CDK2 and DHFR, the CRANkS scores consistently prioritise more unique scaffolds in the top 100 molecules than the fingerprint scores.

For CDK2 and HSP90A, the AUCs of the CRANkS scores and Morgan Fingerprint Scores are similar. Consequently, the scaffold diversity of the methods can be compared more confidently for these targets without considering the effect of active/inactive differentiation. For both these targets for both scaffold diversity measurements the CRANkS scoring ranking outperforms both fingerprinting methods.

Figure 3.21. Table showing the number of unique scaffolds in the top 100 ranked molecules. This is calculated for each target for different methods. For the CRANkS scores and Morgan and 3D Pharmacophore Fingerprint Scores the values are split by the number of input structures, and each calculated value is an average over five separate runs of different input structures. The highest number of scaffolds for each number of input structures is shown in bold, and the highest number of scaffolds across all numbers of input structures is shown with a black outline.

These results indicate that the CRANkS algorithm facilitates scaffold-hopping by prioritising unique scaffolds. The results are target-dependent but generally the CRANkS scores prioritise more unique scaffolds earlier in the ranking than the fingerprint scoring methods by both scaffold diversity calculations used here. However, there is also the issue that scaffold diversity may be inherently linked to the ability to separate actives and inactives. If the actives are prioritised well, then it is likely fewer unique scaffolds are prioritised. However, if the molecules are scored randomly, it is likely that more unique scaffolds will be prioritised. Reassuringly when the AUCs of the scoring methods are similar (as for targets CDK2 and HSP90A), the CRANkS scores prioritise more unique scaffolds.

The diversity of the scaffolds selected can be visualised using an adaption of the tree map view from Scaffold Hunter. The tree map is shown in Figure 3.22. Each small square corresponds to a molecule and each surrounding box corresponds to a scaffold. The figure colours each molecule selected in the top-100 ranked molecules by the scoring method. The figure shows that the Interaction Score, coloured in pink, selects molecules that cover a wider range of scaffold space than the Morgan Fingerprint Score (blue) or 3D Pharmacophore Fingerprint Score (green).

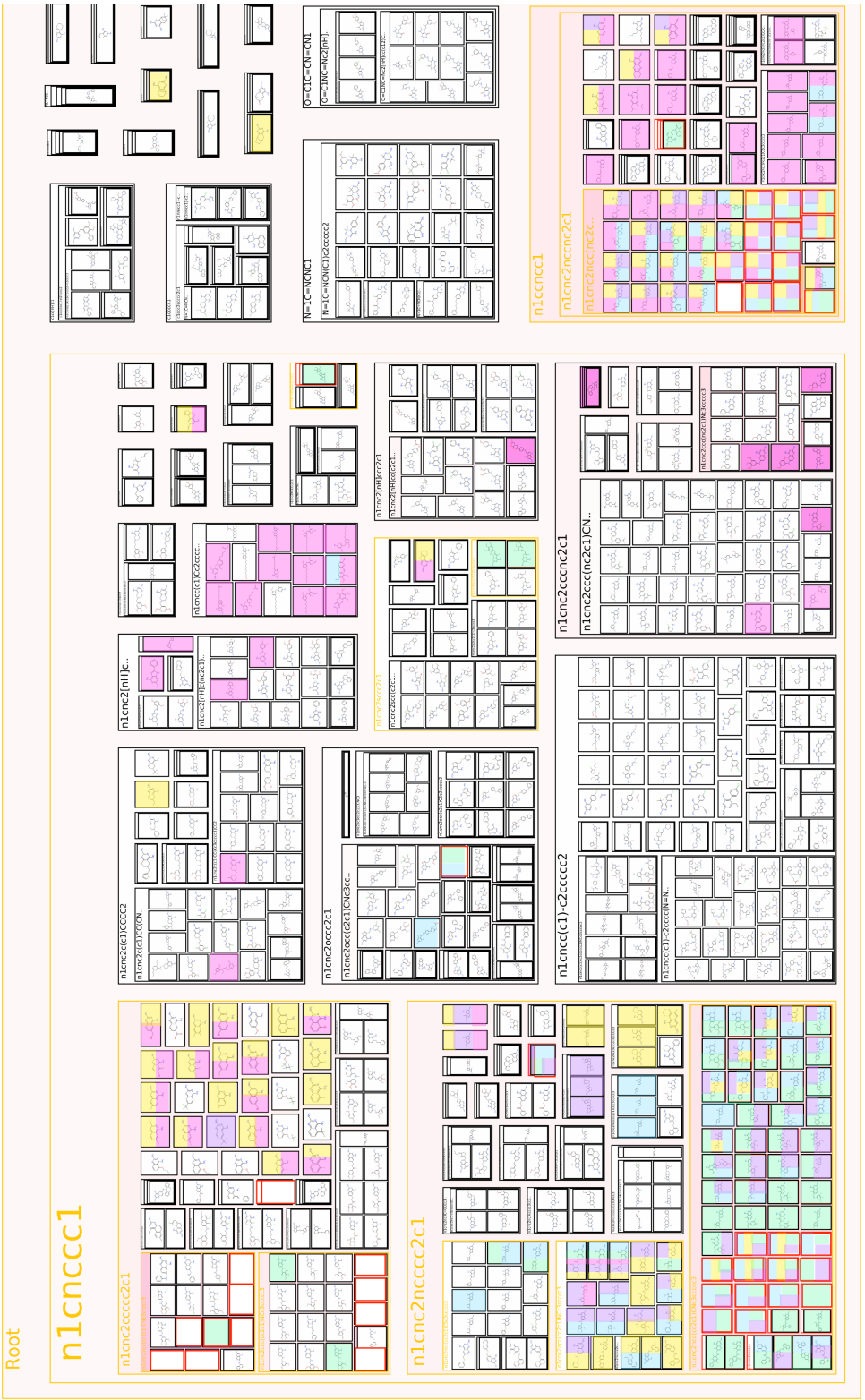



Figure 3.22. Scaffold Tree for candidate compounds scored for DHFR. Active compounds are shown with a red outline. Each box refers to a different scaffold. The top 100 molecules selected by each scoring method are shown, green for the 3D Pharmacophore Fingerprint Score, blue for the Morgan Fingerprint Score, pink for the CRANkS Interaction Score, yellow for the CRANkS Element Score and purple for the CRANkS Pharmacophore Score. The Interaction Score selects compounds over a wider range of scaffold space than the Morgan Fingerprint Score.

### *3.3.2.2 Multi-Objective Optimisation on CRANkS Interaction and Element Score*

Multi-objective optimisation is inherently important in drug discovery, as drug discovery itself is a difficult multi-objective optimisation problem (Nicolau and Brown, 2013). A multi-objective problem contains at least two primary variables which the solutions must satisfy. The values of the two variables can be conflicting which causes a challenge in solving these problems. A multi- objective problem can have multiple equivalent solutions - Pareto solutions (Pareto, 1896).

Originally multi-objective problems were solved by simply optimising a single variable, for example in drug discovery only potency would be optimised followed by the optimisation of solubility for example, *etc*. However, addressing each variable sequentially takes much longer and often results in failure, as early decisions mean objectives addressed later, such as a compound's pharmacokinetics, cannot be met sufficiently (Barignhaus and Matter, 2004). Over the last two decades, there has been a shift to consider the parameters required for a successful drug in parallel. Multi-objective methodologies have been developed that are now commonly-used and have had success in drug campaigns (Lusher *et al.*, 2011; Nicolotti *et al.,* 2002).

In Section 2.2.6.4, the interpretation of the CRANkS novelty scores are discussed. I hypothesise that by maximising the Element Score and minimising the Interaction Score scaffold-hopping can be prioritised. The compounds that satisfy these constraints should form interactions that are similar to those formed by the ligands that make up the grids but also should be novel in terms of the atoms that make up

the compounds, and the placement of those atoms in the binding site. This is effectively a multi-objective optimisation problem. The Pareto optimal solutions to maximising the Element Score and minimising the Interaction Score could be potential candidates for synthesis in a drug discovery campaign if scaffold-hopping is a priority.

To investigate this, Pareto optimal solutions were calculated for each run of the CRANkS algorithm for each dataset. This was implemented using Platypus (Haka, 2015) in Python. The Non- dominated Sorting Genetic Algorithm II (NSGA-II; Deb et al., 2002) was used as it is a fast non-dominated approach, using 10,000 function evaluations. An example of the Pareto solutions found using this method for a run of the BACE1 dataset are shown in Figure 3.23.

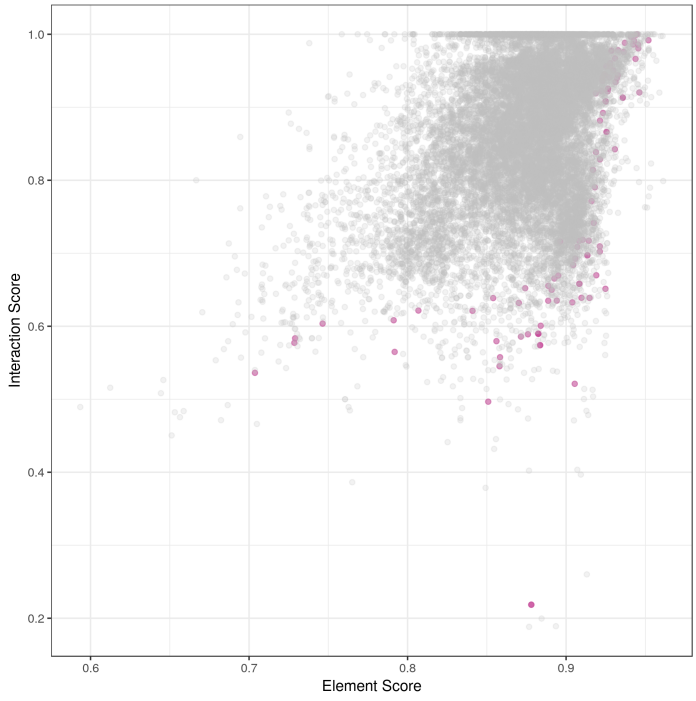



Figure 3.23 The Interaction Score is plotted with respect to the Element Score for all compounds in the BACE1 dataset for the second run of 25 protein-ligand crystal structures. Each compound is shown by a grey point. The calculated Pareto solutions is coloured by pink dots.

The  ability of this process to find active compounds was tested. Figure 3.24 shows a heat map of the percentage of the compounds found to be Pareto solutions that were active. This is compared to the percentage of active compounds in the whole set. The results vary. For the targets ADRB2, ANDR and BACE1 there are consistently good results, with only a few runs that achieve lower active enrichment than the dataset as a whole. For all these targets the Interaction Score and Element Score both achieved high performance in discrimination of actives from inactives across all metrics. The worst performance is achieved for targets PYGM and CDK2 when frequently no active compounds are retrieved. This corresponds to a very poor performance of the Interaction Score for active versus inactive discrimination across all the metrics calculated here. The inabilities of the Interaction Score to prioritise active compounds is likely to be affecting the multi-objective solutions, causing no active compounds to be found. Interestingly, although for the HSP90A datasets all the CRANkS scores achieved AUCs above 0.80, the enrichment of actives in the Pareto solutions is variable. For 50 structures, four of the runs did not recover any active compounds. This could be because both the Interaction Score and Element Score achieve such a high discrimination of actives and inactives, that by maximising one and minimising the other, the maximising of the Element Score is dominating the solution, causing no actives to be recovered.

Figure 3.24 The percentage of molecules that are active when selected as optimal solutions for the multi-objective optimisation of maximising the Element Score whilst minimising the Interaction Score. Enrichment that is higher than that of the whole dataset is shown in pink, whilst enrichment that is lower is shown in blue. Values are shown for each different run for each set of protein-ligand crystal structures used in the CRANkS grids.

To investigate the potential of using this method for scaffold-hopping Figure 3.25 shows the number of unique scaffolds for the compounds calculated to be Pareto solutions. This is compared to the number of unique scaffolds that would be selected at random for the same number of compounds based on the number of unique scaffolds in the whole set. The results are extremely promising, with the number of scaffolds retrieved significantly higher than the estimate based on the distribution of the dataset. This is consistent across all targets, for some runs achieving ten times the number of unique scaffolds, than the estimate based on the distribution of the whole dataset.

Finally, the number of unique active scaffolds in the compounds calculated to be Pareto solutions are shown in Figure 3.26. Based on the number of unique active scaffolds the number of scaffolds that would be retrieved for the same number of compounds, based on the whole dataset, is zero for all targets aside from BRD1. For nearly all targets the number of active scaffolds is highly enriched. The only runs for which the number of active scaffolds is not significantly higher is for four of the runs for HSP90A using 50 protein-ligand structures. Again, this is because the solutions are being dominated by the maximising of the Element Score. The Element Score achieves very high discrimination of actives from inactives when compounds are ranked by the score from low to high. Therfore, if the maximised Element Score was dominating the solution then only inactive compounds would be retrieved.

Figure 3.25 The number of unique scaffolds when compounds are selected as optimal solutions for the multi-objective optimisation of maximising the Element Score whilst minimising the Interaction Score. The number of scaffolds that would be picked based on the distribution of the whole dataset is shown. The pinker the square the higher the number of scaffolds picked. Values are shown for each different run for each set of protein-ligand crystal structures used in the CRANkS grids.

Figure 3.26 The number of unique active scaffolds when compounds are selected as optimal solutions for the multi- objective optimisation of maximising the Element Score whilst minimising the Interaction Score. The number of active scaffolds that would be picked based on the distribution of the whole dataset is shown. The pinker the square the higher the number of active scaffolds picked. Values are shown for each different run for each set of protein-ligand crystal structures used in the CRANkS grids.

These results indicate that treating the selection of compounds as a multi-objective optimisation problem could be beneficial. Maximising the Element Score whilst minimising the Interaction Score was found to enrich for unique active scaffolds across all targets, which could facilitate scaffold-hopping. Further work in this area will include tuning the parameters used in finding the Pareto solutions and investigating why some runs achieved high enrichment for active compounds and some did not.

## 3.4 Conclusions

The CRANkS algorithm has been tested on a variety of datasets in this Chapter. By using benchmarking sets the performance of the CRANkS algorithm at discriminating active compounds and inactive compounds has been investigated and compared to other virtual screening tools. The CRANkS algorithm was found to be target- dependent but overall perform as well as popular commercial docking tools. Notably, the algorithm performed well in initial enrichment and was found to outperform 3 out of 4 commercial docking tools using the BEDROC metric for the majority of targets. The effect of ligand alignment and protein conformations of the input protein-ligand structures on the performance of the algorithm was explored, and this suggested that both ligands that had no overlap in shape, and complexes with differing protein conformations, could be detrimental to the performance of CRANkS. Further work could confirm this by running the CRANkS algorithm on clustered protein-ligand structures. Runs would be performed with structures clustered by the RMSD of the binding site of the protein structures to determine if

this improves performance. This could also be repeated with structures clustered by the shape similarity of the ligands.

The scaffold-hopping potential of the algorithm was also investigated by calculating the number of molecules required to retrieve 50% of the unique scaffolds. The results are target-dependent but generally the CRANkS scores prioritise more unique scaffolds earlier in the ranking than the fingerprint scoring methods by both scaffold diversity calculations used in this work. However, there is also the issue that scaffold diversity may be inherently linked to the separation of actives and inactives. If the actives are prioritised well, then it is likely fewer unique scaffolds are prioritised. However, if the molecules are scored randomly, it is likely that more unique scaffolds will be prioritised. Reassuringly when the AUCs of the scoring methods are similar (as for targets CDK2 and HSP90A), CRANkS prioritises more unique scaffolds. Further work should compare the unique scaffold enrichment to commercial docking tools rather than only comparing to similarity methods.

The ability of the CRANkS algorithm to be used for scaffold-hopping was also investigated using multi-objective optimisation to maximise the Element Score and minimise the Interaction Score. It was hypothesised that the Pareto optimal solutions to this optimisation could be taken forward for synthesis in a drug discovery campaign. These solutions were calculated for each run of the CRANkS algorithm for each target. The enrichment of activity, unique scaffolds and active scaffolds were investigated. Although the enrichment of active compounds varied between runs, across all targets, for nearly all runs there was significant enrichment of unique scaffolds and of active scaffolds. These results indicate that treating the selection of

compounds as a multi- objective optimisation problem could be beneficial. Further work in this area should include tuning the parameters used in finding the Pareto solutions and investigating why some runs achieved high enrichment for active compounds and some did not.

Overall, I have shown the CRANkS algorithm prioritises active novel scaffolds using a data-driven approach. The algorithm is shown to have comparable ability to leading commercial docking tools in separating actives from inactives without the need for a complicated protocol. However, the real benefit to the algorithm comes from its ability to facilitate scaffold-hopping. I show that the algorithm prioritises more unique scaffolds maintaining a higher scaffold diversity, and consequently better coverage of chemical space. Additionally, I have developed a method for selecting compounds for follow-up chemistry using multi-objective optimisation to maximise the Element Score and minimise the Interaction Score to prioritise novel active scaffolds.

Further work, in addition to the testing outlined above, would primarily involve improvement of the conformer generation. This was found to still frequently produce conformations with an RMSD of up to 20 Å away from the original crystal structure. This inaccuracy is likely to be detrimental to the performance of the algorithm - if the conformers calculated are very different to the actual conformations and position in the binding site adopted by the compound, the protein-ligand interactions calculated will not be correct, and the overlap with the CRANkS grids will not be useful information. Thus, the algorithm should be modified to make use of other conformation generation tools, such as docking, with a

higher proven accuracy. This has been investigated by combining a modified version of CRANkS with AutoDock in Chapter 4, and modifying the algorithm to use constrained docking in Chapter 5.

# Chapter 4

# Active-Guided Docking

## 4.1 Introduction

In the realm of structure-based virtual screening, protein-ligand docking has become one of the most widely used scoring methods for lead identification (Sousa *et al.,* 2013). It has been shown to be capable of finding sub micro-molar inhibitors (Ripphausen *et al.,* 2012). There is a consensus that docking methods can predict binding modes relatively well – often with an RMSD of within 2 Å of a crystal structure (Wang *et al.,* 2016). However, there is still a long way to go in accurately predicting the affinity of a ligand for a protein (Moitessier *et al.,* 2008), and even pose prediction can be improved for ligands with novel scaffolds (Gaieb *et al.,* 2018).

Docking methods tend not to make use of experimental data on the target of interest, such as knowledge about the structures and binding modes of ligands that bind. The only data currently used is in the training of scoring functions, and this is often on a variety of targets and not necessarily the target or target class of interest - historically this data has not been available. It is well-known that the performance of scoring functions is inherently target-dependent and "one size does not fit all" (Ross *et al.,* 2013).

In this Chapter I combine this structural data with AutoDock 4, the most highly cited open-source docking tool (Morris *et al.,* 2009; Morris *et al.,* 1998). AutoDock works by making use of AutoGrid which pre-calculates interaction potentials for the different atom types in the ligand, using a single protein structure, as well as desolvation and electrostatic potential maps. AutoDock then uses these maps to generate binding poses and score the poses. I combine these grid maps with a slightly modified version of the CRANkS grids described in Chapters 2 and 3.

By making use of the structural data already available for active bound ligands a target-dependent improvement on both binding prediction and scoring performance was achieved. Different weightings of combining CRANkS grids and AutoGrid maps were explored using a smaller number of targets from the DEKOIS 2.0 dataset (Bauer *et al.,* 2013), and the results here indicate the parameters that should be used on a much larger testing of the method. Two different types of modified CRANkS grids were tested. AutoCRANkS combines the signal for all atoms in each of the ligands of the structural data with the AutoGrid maps. AutoCRANkS Int only combines the signal from atoms of ligands calculated to form interactions with the protein target with AutoGrid maps. This interaction-filtering was found to improve the performance of AutoDock but to a lesser extent than for AutoCRANkS.

## 4.2 Method

### *4.2.1 Combination of CRANkS Grids with AutoGrid maps*

AutoDock uses a set of grid maps generated by AutoGrid to dock and score compounds (Morris *et al.,* 1998). The AutoGrid maps used by AutoDock are a set of atom affinities at each grid point for each AutoDock atom type. The maps have a default spacing of 0.375 Å. In Chapter 2 I introduced the CRANkS grids which describe the structural data of protein-ligand complexes of the target of interest. There are three types of grids – the element grid which describes the elements of each atom in each active ligand, the pharmacophore grid which describes the pharmacophoric features of each ligand and the interaction grid which describes the protein-ligand interactions of each ligand. I decided to combine the grids in two ways. The first makes use of the element and pharmacophoric features of each atom to assign an AutoDock atom type to each atom in the ligand. Signals from each of these atoms can then be added to the corresponding AutoGrid map. We refer to this method as AutoCRANkS. The second only adds the signal from an atom if the atom forms part of a protein-ligand interaction. We refer to this method as interaction-filtered AutoCRANkS (AutoCRANkS Int). Both ways of combining the maps were tested to explore whether filtering the atoms by interactions could reduce noise and improve results. The details of the map combinations are discussed below.

### *4.2.1.1 Mapping to AutoDock Atom Types.*

Table 4.1 details the conversion from the elements and pharmacophoric features calculated by RDKit for an atom, to a corresponding AutoDock type. Combinations of elements and features, or conversely lack of features, indicate properties about the atom and allow the assignment. These were determined using the description of each of the AutoDock atom types.

| Element | | RDKit Pharmacophoric Feature | | AutoDock Atom Type | AutoDock Description |
|---|---|---|---|---|---|
| Carbon | + | Aromatic | = | A | Non H-bonding Aromatic Carbon |
| Carbon | + | Not Aromatic | = | C | Non H-bonding Aliphatic Carbon |
| Hydrogen | + | Donor | = | HD | Donor 1 H-bond Hydrogen |
| Hydrogen | + | Not Donor | = | H | Non H-bonding Hydrogen |
| Nitrogen | + | Not Acceptor | = | N | Non H-bonding Nitrogen |
| Nitrogen | + | Acceptor | = | NA | Acceptor 1 H-bond Nitrogen |
| Oxygen | + | Acceptor | = | OA | Acceptor 2 H-bonds Oxygen |
| Oxygen | + | Not Acceptor | = | OS | Acceptor S Spherical Oxygen |
| Phosphorus | + | Any | = | P | Non H-bonding Phosphorus |
| Sulphur | + | Acceptor | = | SA | Acceptor 2 H-bonds Sulphur |
| Sulphur | + | Not Acceptor | = | S | Non H-bonding Sulphur |
| Fluorine | + | Any | = | F | Non H-bonding Fluorine |
| Chlorine | + | Any | = | Cl | Non H-bonding Chlorine |
| Bromine | + | Any | = | Br | Non H-bonding Bromine |
| Iodine | + | Any | = | I | Non H-bonding Iodine |

Table 4.1. Conversion of RDKit pharmacophoric features and elements to AutoDock atom types.

### *4.2.1.2 Addition of CRANkS Signal to AutoGrid maps.*

AutoGrid maps for each of the atom types described in Table 4.1 were generated with a spacing of 0.375 Å. To generate the AutoCRANkS maps for each of the ligands to be used in the grid, each atom is taken in turn. The atom is assigned an AutoDock atom type and the nearest grid point to the atom is found. A spherical Gaussian is then applied with a standard deviation of half the van der Waals radius of the given atom type. This follows the work of Anighoro and Bajorath who successfully used Gaussian distributions to calculate 3D similarity of actives and decoys to ligands with promising results in terms of active-decoy discrimination (Anighoro and Bajorath, 2016).

The functional form of the AutoDock scoring function is shown in Equation 4.1 (Morris *et al.,* 1998):

$$\Delta G = \Delta G_{vdW} \sum_{i,j} \left( \frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}} \right) + \Delta G_{hbond} \sum_{i,j} E(t) \left( \frac{C_{ij}}{r_{ij}^{12}} - \frac{D_{ij}}{r_{ij}^{10}} + E_{hbond} \right) + \Delta G_{elec} \sum_{i,j} \frac{q_i q_j}{\varepsilon(r_{ij}) r_{ij}} + \Delta G_{tor} N_{tor} + \Delta G_{sol} \sum_{i_C,j} S_i V_j e^{(-r_{ij}^2/2\sigma^2)} \quad (4.1)$$

Each of the five ΔG terms are coefficients that were determined using linear regression. The scoring function can be split into five energetic contributions. The first is a Lennard-Jones repulsion term. The second is a hydrogen-bonding term which uses *E(t)* to weight the term using the angle *t* between the protein and ligand atom. Third is the electrostatic contribution calculated using a Coulombic potential. Fourth is a term to measure entropy using $N_{tor}$: the number of $sp^3$ bonds in the

ligand. The final term measures the desolvation energy to capitulate desolvation on binding and the hydrophobic effect. These are measured over all pairs of ligand atoms, *i* and protein atoms, *j*.

AutoDock uses AutoGrid to pre-calculate a number of these terms at each grid point for each AutoDock atom type. An electrostatic map and desolvation map are generated that contain the electrostatic and desolvation potential at each grid point respectively. A map is also generated for each AutoDock atom type containing the sum of the van der Waals and hydrogen bond contributions. These are the maps that are modified by AutoCRANkS.

A depiction of the addition of signal from ligand atoms to the AutoGrid atomic affinity maps is shown in Figure 4.1. A gaussian is applied centred on the ligand atom coordinates. This signal is multiplied by a normalisation constant *w*. This contribution is then added to the atomic affinity at the corresponding grid points in the atomic affinity map of the same AutoDock atom type. The functional form of the atomic affinity at a given grid point can be written as:

$$\Delta G = \frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}} + E(t)\left(\frac{C_{ij}}{r_{ij}^{12}} - \frac{D_{ij}}{r_{ij}^{10}} + E_{hbond}\right) - \sum_a \frac{w}{\sigma_{vdw}^3 (2\pi)^{3/2}} \exp\left(-\frac{(x_i - x_a)^2 + (y_i - y_a)^2 + (z_i - z_a)^2}{\sigma_{vdw}^2}\right) \quad (4.2)$$

This combines the Van der Waals and hydrogen bond contributions with a 3-dimensional gaussian distribution summed over atoms from active ligands, *a*. The mean of the distribution is the coordinates of *a* and the standard deviation $\sigma_{vdw}$ is half

the Van der Waals radius of the element of the AutoDock atom type. The gaussian is weighted by the normalisation constant *w*.

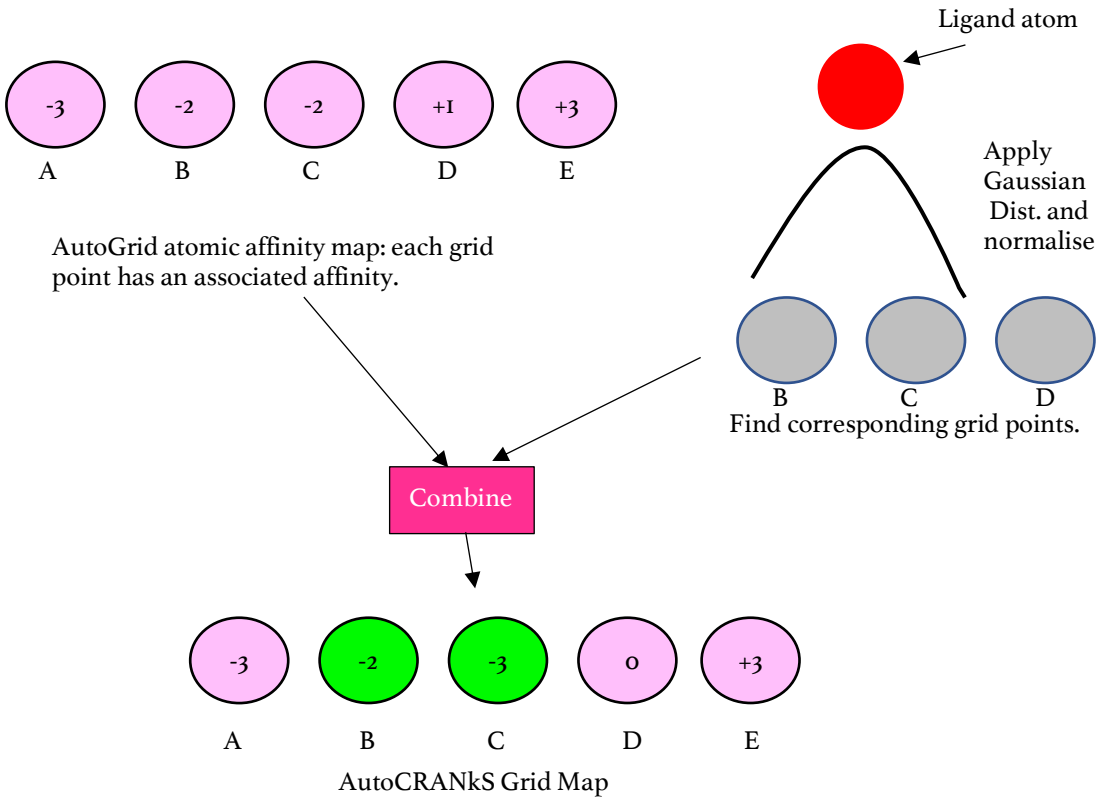



Figure 4.1. Depiction of the generation of AutoCRANkS grid maps. For each ligand atom a gaussian is applied at the coordinates of the atom and this is normalised by a normalisation constant. The grid points covered by the distribution are located and the value of the distribution at that grid point is taken away from the atomic affinity i.e. the number is made more negative and therefore the potential more attractive.

In order to determine the ideal contribution of the CRANkS maps, five weights or normalisation constants were applied to the Gaussian values: 0.01, 0.1, 1, 10 and 100. These normalised values were then subtracted from the corresponding value of the AutoGrid map at the same grid point. This makes the AutoGrid map more attractive at grid points that overlap with a Gaussian centred on a ligand atom.

For AutoCRANkS Interaction-filtered grid maps, signal is only added to the AutoGrid maps if the atom is calculated to form an interaction with the protein. Interactions are calculated using the command-line version of PLIP as described in Chapter , Section 3.2.2.2.

### *4.2.2 Datasets*

The AutoCRANkS and AutoCRANkS Int algorithms were tested on sets of actives and decoys for a variety of targets. For this work targets were selected from the DEKOIS 2.0 dataset (Bauer *et al.,* 2013). This dataset contains 40 active compounds and 1200 decoys each for 81 diverse targets. This dataset allows comparison to the performance of AutoCRANkS and AutoCRANkS Int to the popular docking tool AutoDock Vina (Trott *et al.,* 2010; Bauer *et al.,* 2013), as well as commercial tools GOLD (Jones *et al.,* 1997; Bauer *et al.,* 2013) and Glide (Friesner *et al.,* 2004; Bauer *et al.,* 2013). The DEKOIS 2.0 dataset consists of significantly fewer actives and decoys for each target than for the DUD-E dataset used in Chapter 3. This reduces the computational time required to test an individual target.

The dataset was split into six drug target classes: chaperone proteins (1 target), enzymes (33 targets), kinases (22 targets), receptors (15 targets), regulators (1 target) and miscellaneous proteins which did not fit into any other category (9 targets). A target was selected from each class. Table 4.2 details the six targets selected.

| Target Name | Abbrev-iation | Drug Target Class | UniProt Code | Number of PDB Structures (UniProt) | Number of PDB Structures After Inspection |
|---|---|---|---|---|---|
| Heat shock protein 90 alpha | HSP90A | Chaperone | P07900 | 76 | 74 |
| β-2 adrenergic receptor | ADRB2 | Receptor | P07550 | 7 | 7 |
| Dihydrofolate reductase | DHFR | Enzyme | P00374 | 65 | 63 |
| Kinesin family member 11 | KIF11 | Motor | P52732 | 29 | 29 |
| Apoptosis regulator B-cell lymphoma 2 | BCL2 | Regulator | P10415 | 16 | 9 |
| RAC-α serine/threonine -protein kinase (PKB) | AKT1 | Kinase | P31749 | 12 | 12 |

Table 4.2. Details of the six targets from the DEKOIS 2.0 dataset used to test AutoCRANkS and AutoCRANkS Int.

For each class, if there was a target that had been tested in Chapters 2 and 3 this target was selected as it would allow comparison to results found in that work. Otherwise targets were selected that had a high number of PDB structures. The PDB structures for each target were collected using the UniProt code for the target. The structures were then aligned using PyMOL (Schrödinger, LLC) and visually inspected. Any structures with ligands identical to structures already in the set, ligands that were not in the binding site, or no ligands were removed. The full set of PDB structures and reasons for any rejections for the set can be found in Appendix B, Table B.3.1 for ADRB2, Table B.3.5 for HSP90A, Table B.3.8 for DHFR and Tables B.4.1 to B.4.3 for targets KIF11, BCL2 and AKT1 respectively.

### *4.2.3. Testing*

For each target, five grids were tested. For each of these grids the AutoCRANkS and AutoCRANkS Int algorithms were run for the five normalisation constants: 0.01, 0.1, 1, 10, 100. To help determine the effect of AutoCRANkS on binding pose generation and scoring function separately, the poses generated by AutoCRANkS were scored by AutoDock (refered to as AutoDock | AutoCRANkS). The poses generated by AutoDock were also scored by AutoCRANkS (refered to as AutoCRANkS | AutoDock). By comparison of these results to the results for AutoCRANkS I can explore the effect of changing the poses or scoring and better determine how the CRANKs grids are changing the protocol.

#### *4.2.3.1 Active Grids*

The PDB structures used in each of the grids can be found in Appendix B, Table B.4.4. The three letter residue codes for each ligand are shown in Appendix B, Tables B.3.12 to B.3.19 for ADRB2, Tables 3.59 to 3.63 for HSP90A, Tables 3.118 to B.3.122 for DHFR and Tables B.4.1 to B.4.3 for KIF11, BCL2 and AKT1 respectively. For targets HSP90A, DHFR and ADRB2 the same ligands that were used for each grid in Chapter 3 are used again here. These ligands were randomly selected in Chapter 3.

To determine the effect of the choice of ligands on the docking results, for the remaining targets the ligands were clustered using Morgan Fingerprints and

Tanimoto similarity (see Chapter 3, Section 3.2.5.1). This clusters based on molecular similarity and could indicate how ligands should be selected in future experiments i.e. whether a diverse set should be used, or ligands within one cluster. A cut-off similarity of 0.6 was used.

For AKT1 clustering using Morgan Fingerprints showed that 4 out of the 12 ligands from the PDB structures clustered into a single cluster. All other ligands were calculated to be in separate clusters. Grid 1 uses the four ligands found to be in a single cluster. This grid represents choosing similar ligands to construct the CRANkS grids. Grids 2, 3 and 4 used one ligand from the cluster and 3 additionally randomly selected ligands from the individual clusters. Grid 5 only uses ligands from individual clusters. These sets indicate a diverse selection of ligands to build the grids.

For target BCL2 the nine ligands were clustered using Morgan Fingerprints into one large cluster of seven ligands. The remaining two ligands did not form a cluster. Grids 1, 2 and 3 contain five ligands only from the cluster. These grids are constructed from molecularly similar ligands. Grids 4 and 5 use ligands picked randomly from all possible nine ligands and are consequently constructed from a diverse set of ligands.

For target KIF11 all ligands were found to form one cluster when clustered using Morgan Fingerprints. Consequently five grids were selected consisting of five randomly selected ligands from the cluster. All of these grids can be considered to be constructed from molecularly similar ligands.

### *4.2.4 Docking Protocol*

For each target an identical docking protocol was adopted. Figure 4.2 shows a visual depiction of the protocol. The treatment of proteins and ligands was adapted from work by Susan Leung (*private communication*). The PDB structure used to dock for each target was the structure specified in the DEKOIS 2.0 dataset and are given in Appendix B, Table B.4.5. For each structure the number of points in the $x$, $y$ and $z$ direction for a grid of 0.375 Å and the centre of the grid were determined. These were calculated by visual inspection by ensuring all ligands from all PDB structures for a given target were covered by the grid points (Appendix B, Table B.4.5).

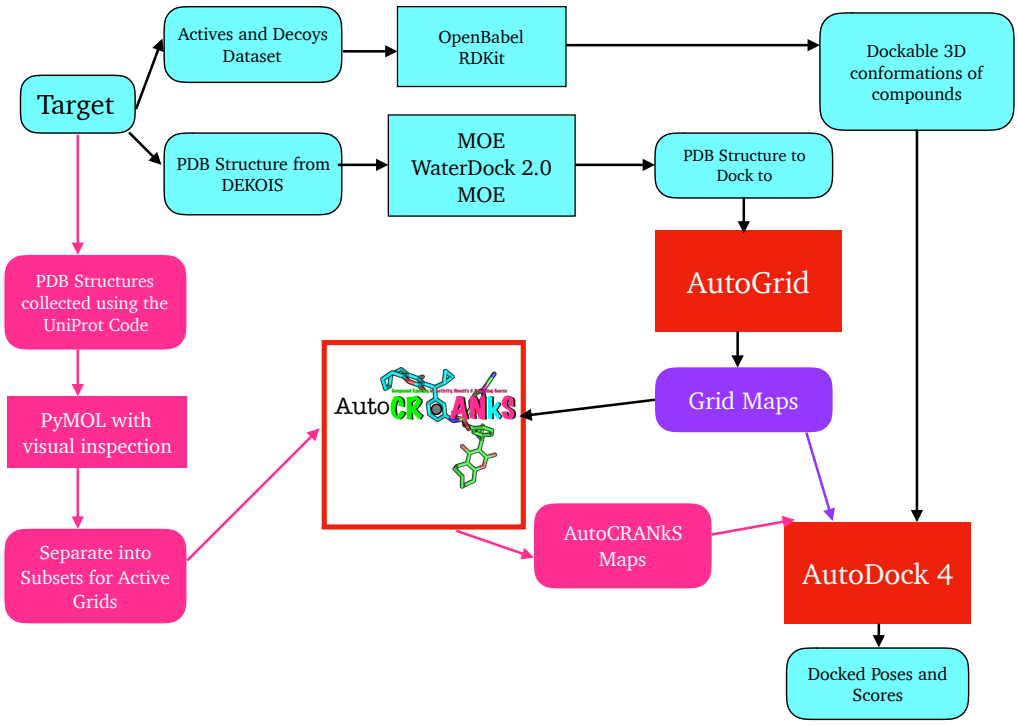



Figure 4.2. Docking protocol for AutoCRANkS and AutoDock for the targets from the DEKOIS 2.0 dataset tested as part of this work. Steps that are part of both protocols are shown with black arrows, steps that are only in the AutoCRANkS protocol are shown by pink arrows and steps that are only in the AutoDock protocol are shown by purple arrows.

First the ligand is removed from each of the PDB structures to be docked to. The protein structures were then protonated using the *Protonate3D* functionality of MOE to assign partial charges (Molecular Operating Environment MOE, 2013). Water molecules can have a large effect on the results of docking (as discussed in Chapter 1, Section 1.4.3) and water molecules will not be displaced when using AutoDock. Consequently, to ensure only conserved waters are using in the docking rather than all waters in the structure, WaterDock 2.0 was applied to each structure (Ross *et al.,* 2012, Sridhar *et al.,* 2017). By docking water molecules using AutoDock Vina (Trott *et al.,* 2010), WaterDock 2.0 predicts the position of water molecules in a protein structure. For the resulting docked water molecules, if any water molecules from the original PDB structure are within 2 Å of the docked water, the water molecules are kept in the structure. Otherwise the water molecules are removed. This typically leaves 2-4 water molecules in the structure but can remove all waters. The resulting protein structure was then protonated again using the *Protonate3D* functionality of MOE.

For the actives and decoys of each of the DEKOIS 2.0 datasets, a 3D conformation was generated using the "*ETKDG*" method from RDKit (Landrum, 2015). This uses distance geometry to generate conformers but uses torsion angle data from the Cambridge Structural Database to amend the conformations (Riniker and Landrum, 2015). These conformations were then protonated at pH 7 using OpenBabel (O'Boyle *et al.,* 2011). For redocked ligands as described in Section 4.3.1.1 the 3D conformation of the ligand was unchanged. The ligands were also similarly protonated using OpenBabel.

AutoGrid was then used to generate grid maps for each target. Grid maps were generated for the atom types listed in Table 4.1. The AutoGrid maps were modified for AutoCRANkS and AutoCRANkS Int using Python and Numpy. Each ligand to be used in the active grid was protonated using OpenBabel and handled by RDKit. For each atom the nearest grid point was found using the *"kd-tree"* functionality in Scipy (Maneewongvatana and Mount, 1999, Jones *et al.,* 2001). Given the normalisation constant the grid is then modified as described in Section 4.2.2.2.

The docking of the actives and decoys can then be run using AutoDock with the corresponding grid maps from AutoGrid, AutoCRANkS or AutoCRANkS Int. A rate of gene mutation of 0.2 and a crossover rate of 0.5 were used with 2,500,000 evaluations of the genetic algorithm. This number of evaluations should be short enough to not be too computationally expensive, but long enough to give adequate solutions. Ten poses were generated for each compound.

## 4.3 Results

### *4.3.1 AutoCRANkS*

#### *4.3.1.1 Re-Docking*

As an initial test to determine whether there is any improvement on incorporating AutoCRANkS grids in the AutoDock docking protocol, the PDB structures were redocked. For all remaining PDB structures for a given target that had not been used

in a given grid, the ligand was redocked using AutoDock and standard AutoGrid maps, followed by AutoCRANkS computed with five different normalisation constants for different grid sets of ligands. For clarity the original AutoDock protocol using AutoGrid maps will be referred to as AutoDock. The results were then analysed in terms of the RMSD of the docked poses, using all atoms, from the original crystal structure. The pose with the closest RMSD to the crystal structure, the RMSD of the pose with the lowest score compared to the crystal structure, and the average crystallographic RMSD over all poses were examined. For each run the percentage of docked structures achieving an RMSD of < 2 Å, and the percentage of docked structures with an RMSD of > 3 Å were calculated.

The number of structures with an RMSD of < 2 Å compared to the crystal structure was calculated to indicate how many structures were redocked to poses close to the original structure and would indicate a successful redocking. The number of structures with an RMSD > 3 Å indicate how many structures were redocked unsuccessfully and represent a failure to redock. Results for a selection of the targets are discussed below.

For target ADRB2 only two structures were able to be redocked as there are only seven PDB structures in total for this target. Thus, it is hard to infer meaning from these results, and there is no significance difference shown between the performance of AutoCRANkS and original AutoDock. The results are shown in Appendix A, Figure 4.1. Similarly. in the case of BCL2 only 4 structures are able to be redocked for each grid as there are only nine structures in total. Therefore, these results should be interpreted with caution. The results are shown in Appendix A, Figure 4.2. There is a

reduction in performance for AutoCRANkS compared to AutoDock for normalisation constants of 1, 10 and 100.). This indicates that the addition of the active grids may be disrupting the scoring function for this target. However, a normalisation constant of 0.1 shows improvement on these metrics indicating that for lower normalisation constants the active grids may add valuable information when the contribution is not dominating.

For the remaining targets the results are promising. There is an improvement in the average RMSD over conformers. This is particularly the case for the RMSD between the lowest scored conformer and the crystallographic structure indicating an improvement in the scoring function when using AutoCRANkS. DHFR and HSP90A are typical of these results and are discussed here. For the remaining targets results can be found in Appendix A, Figures 4.1 to 4.4. The redocking results for DHFR are shown in Figure 4.3. For all grids, apart from grid 4, there is an improvement for all normalisation constants greater than 0.01 over AutoDock. For grid 4, only a normalisation constant of 0.1 exhibits improved performance when using AutoCRANkS.

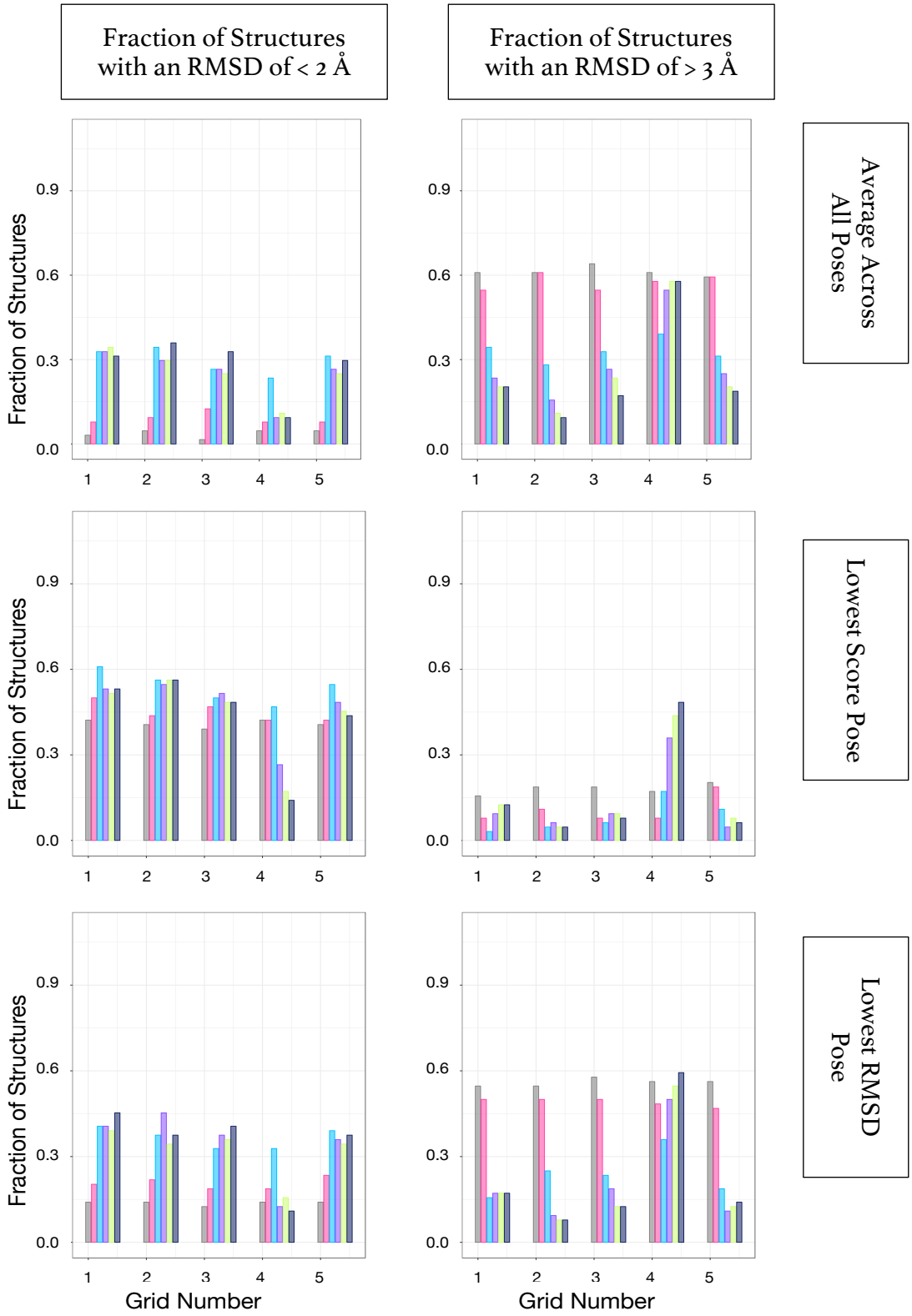



Figure 4.3. Fraction of structures with an RMSD of < 2 Å and an RMSD of > 3 Å are shown for the redocking of PDB structures for DHFR. This is split into the average RMSD between the structure and all conformers, the RMSD between the structure and lowest scored pose, and the lowest RMSD between the structure and any conformer. The normalisation constant is coloured with grey for AutoDock, pink for 0.01, light blue for 0.1, purple for 1, green for 10 and dark blue for 100.

An example of a redocked ligand for DHFR (residue name LII, PDB structure 1kmv) for which AutoCRANkS achieved lower RMSDs between the docked pose and the crystal structure than AutoDock is shown in Figure 4.4. The lowest scored conformer is shown in each case. The corresponding RMSD values are shown in Table 4.3. AutoDock achieves an RMSD for the lowest scored conformer of over 6 Å although the lowest RMSD achieved for any conformer was just over 1 Å. Using AutoCRANkS with a normalisation constant of 0.1 showed a significant improvement of the RMSD between the lowest scored pose and the crystal structure for both grids 1 and 2 (Figure 4.4 A and B, Table 4.3) . The fused aromatic rings are nearly perfectly aligned, and the other aromatic ring is very close to the crystal structure. Both poses show a similar scoring pattern with both rings contributing strongly to the score.

| 1kmv | Lowest Score Conformer: RMSD / Å | Lowest RMSD Conformer: RMSD / Å |
|---|---|---|
| AutoDock | 6.3 | 1.1 |
| Grid 1, $w$=0.1 | 1.2 | 1.1 |
| Grid 1, $w$=10 | 5.5 | 5.2 |
| Grid 2, $w$=0.1 | 1.2 | 1.1 |
| Grid 2, $w$=10 | 1.5 | 1.4 |

Table 4.3. RMSD of the lowest score conformer and the lowest RMSD conformer generated for 1kmv, ligand LII by AutoDock and AutoCRANkS using different grids and normalisation constants.

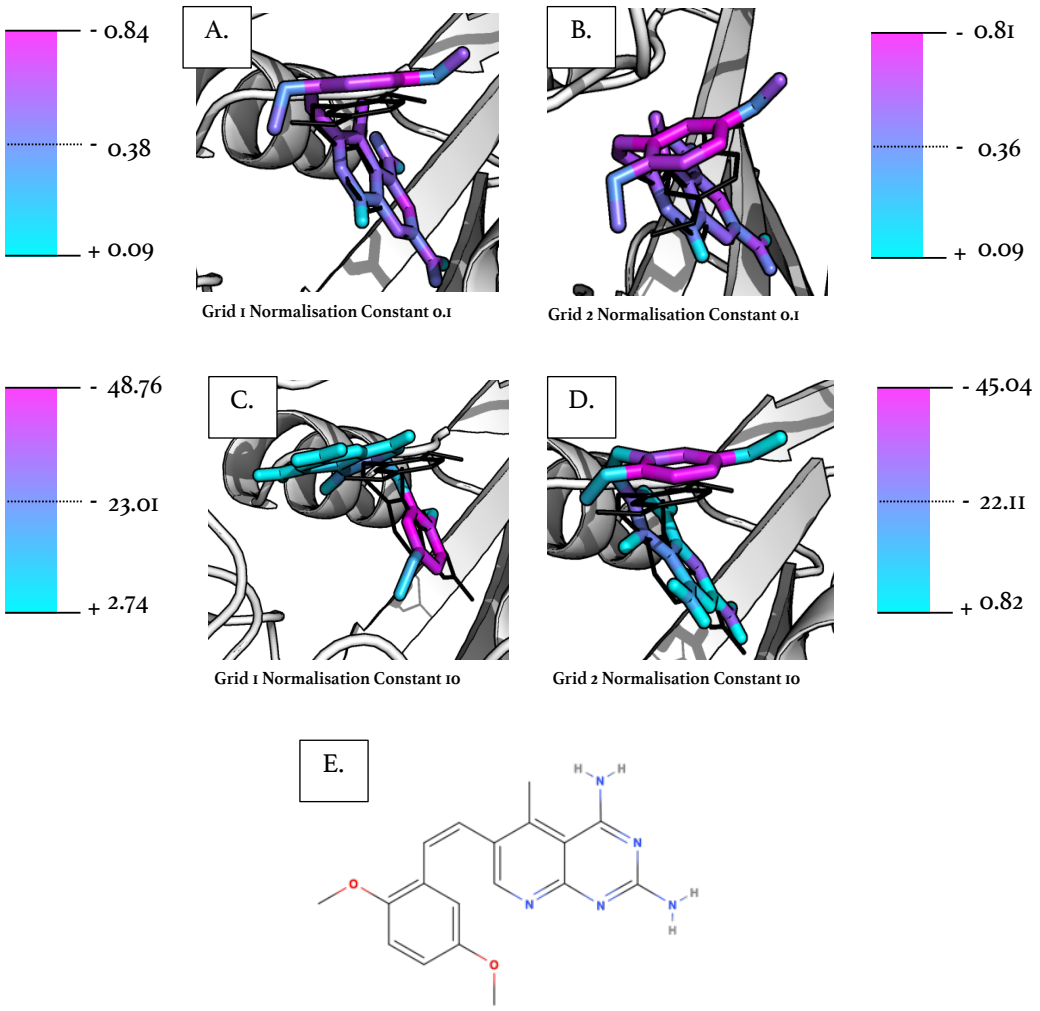



Figure 4.4. Redocked poses for the ligand LII of PDB structure 1kmv using grids 1, normalisation constant 0.1 (A) and normalisation constant 1 (C) and grid 2 with normalisation constants 0.1 (B) and 10 (D) shown as sticks. The X-ray crystallographic ligand structure is shown as a black wireframe. The docked pose is coloured by the per atom contribution to the score. For grid 2 for both normalisation constants a low crystallographic RMSD is achieved. However, for grid 1 a normalisation constant of 0.1 achieves a low crystallographic RMSD whereas when a normalisation constant of 10 is used the aromatic rings have swapped positions yielding a high RMSD. The 2D structure of the ligand is shown in E.

However, as the normalisation constant increases to 10, grid 1 gives a high RMSD of 5.5 Å for the lowest scoring conformer (Figure 4.4 C), while grid 2 continues to achieve a low RMSD (Figure 4.4 D, Table 4.3). Examining the grids, shown in Figure 4.5, show that the pose generated by grid 2 hits two aromatic attractive wells in the aromatic carbon grid map (Figure 4.5 D), whereas the pose generated by grid 1 only overlaps with one (Figure 4.5 A). In addition, the molecule has flipped in the pose

generated by grid 1 so that the rings are in the opposite position. In the case of grid 2 a nitrogen on one of the rings has overlapped with an attractive nitrogen well added by the active grids to the nitrogen acceptor map (Figure 4.5 F). This keeps the rings in the correct position. In contrast grid 1 does not contain this well allowing the rings to swap positions (Figure 4.5 C). This highlights the importance of the weight of contribution of the active grids, but also how the make-up of the grid can affect the performance.

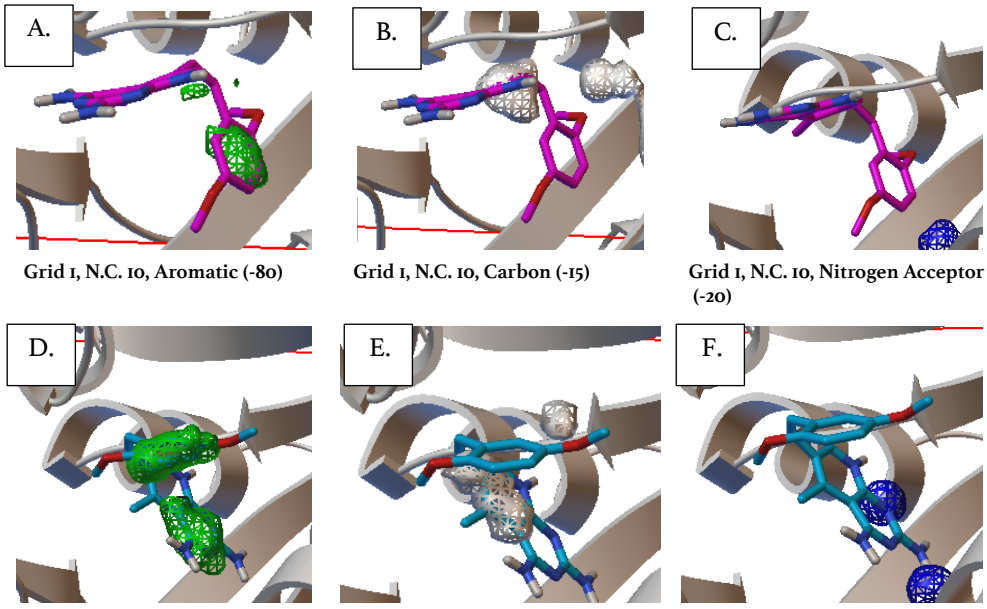



Figure 4.5. AutoCRANkS grid maps for DHFR shown with docked poses for ligand LLI from PDB structure 1kmv using grids 1 and 2 for a normalisation constant of 10. The isocontour level (in kcal/mol) for each map is shown in brackets. The colour of the map is green for the aromatic carbon grid map (A for grid 1, D for grid 2), grey for the aliphatic carbon grid map (B for grid 1, E for grid 2) and blue for the nitrogen acceptor map (C for grid 1 and F for grid 2). The two attractive wells in the aromatic map in grid 2 are covered by the pose as well as overlap with the carbon and nitrogen acceptor wells. In comparison, grid 1 only exhibits one large attractive well in the aromatic map and no overlap with the nitrogen acceptor map.

The results of the redocking using AutoDock and AutoCRANkS for HSP90A are shown in Appendix A, Figure A.4.3. For the average RMSD between all the

conformers and the crystal structure, there is a significant improvement for all normalisation constants greater than 0.01 compared to using AutoDock. There is an increase in the percentage of structures with an average RMSD of < 2 Å and a decrease in the percentage of structures with an average RMSD of > 3 Å, in particular for a normalisation constant of 1 compared to AutoDock. The same behaviour is exhibited when using the RMSD between the lowest score conformer and the crystal structure, indicating that the addition of information about the actives from the AutoCRANkS grids is improving the AutoDock scoring function. The results when using the lowest RMSD of any conformer show an improvement in both measures for a normalisation constant of 1. However, there is no significant improvement for the other normalisation constants.

An example of a redocked structure is shown in Figure 4.6, for the redocking of residue PU3 (9-butyl-8-[(3,4,5-trimethoxyphenyl)methyl]purin-6-amine) from PDB structure 1uy6. The RMSDs between the redocked structure using AutoCRANkS with grid 3 for different normalisation constants and the crystallographic structure is given in Table 4.4. The poses shown are the lowest scored conformer that was generated (Figure 4.6). The table indicates that conformers close to the crystal structure were possible in all cases as the lowest RMSD of any conformer is consistently low (0.84-2.69Å for AutoCRANkS and 0.76Å for AutoDock). However, the RMSD of the scored poses shows a decrease from 3.16Å when using AutoDock to 1.01-1.45Å when using normalisation constants from 0.01 to 1 with AutoCRANkS, indicating that using this weighting of contribution the scoring function is being aided by the CRANkS grid.

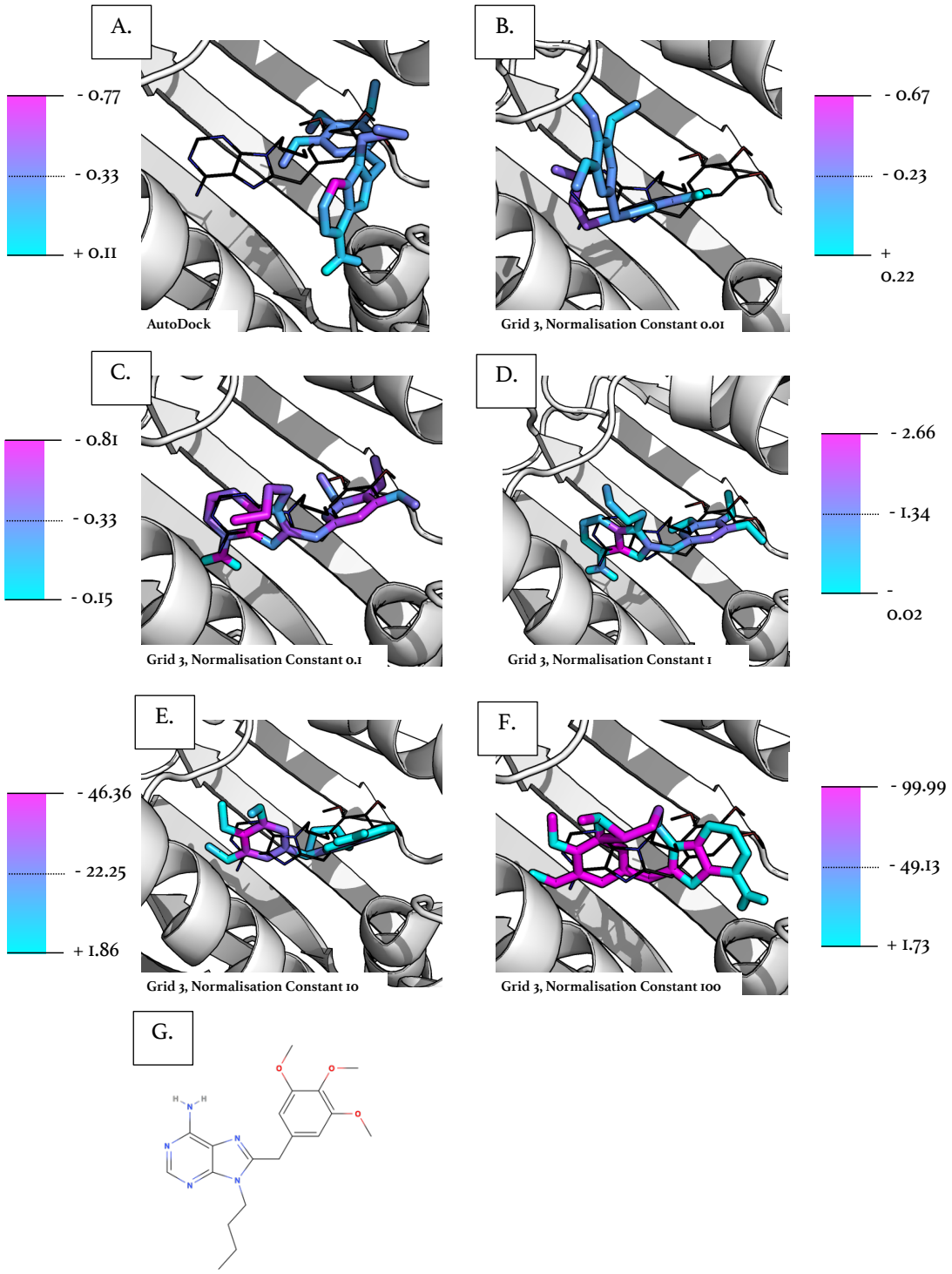
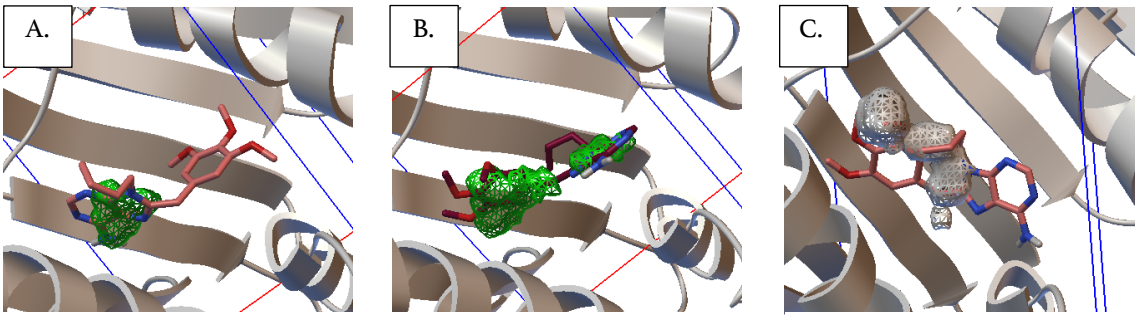



Figure 4.6. Docked poses of ligand PU3 using AutoDock and AutoCRANkS for PDB structure 1uy6. The crystallographic ligand is shown in black wireframe. The poses are coloured by the per atom contribution to the score. Poses are shown generated by AutoDock (A) and AutoCRANkS using grid 3 with a normalisation constant of 0.01 (B), 0.1 (C), 1 (D), 10 (E), and 100 (F). The 2D structure of PU3 is shown in G. The poses closest to the crystallographic structure are found using AutoCRANkS with a normalisation constant of 1 or 0.1.

| Method | Lowest Score Conformer: RMSD / Å | Lowest RMSD Conformer: RMSD / Å |
|---|---|---|
| AutoDock | 3.2 | 0.8 |
| Grid 3, *w*=0.01 | 2.4 | 0.9 |
| Grid 3, *w*=0.1 | 1.1 | 0.8 |
| Grid 3, *w*=1 | 1.5 | 1.2 |
| Grid 3, *w*=10 | 4.6 | 1.5 |
| Grid 3, *w*=100 | 4.6 | 2.7 |

Table 4.4. RMSD of the lowest score conformer and the lowest RMSD conformer generated for residue PU3 of PDB structure 1uy6 by AutoDock and AutoCRANkS using different grids and normalisation constants.

The conformer is coloured by the score per atom (Figure 4.6). In the case of the pose docked by original AutoDock (Figure 4.6 A) there is a high RMSD between the pose and the original crystal structure for the lowest scored conformer (3.2 Å) . The favourably scored atom contribution to the score is the nitrogen of the aromatic 6-membered ring. However, this is not in the same place as the crystal structure. For a normalisation constant of 0.01, a similar pattern is shown with the same atoms achieving similar scores and the RMSD remains high (Figure 4.6 B). When the normalisation constant increases to 0.1 and 1, a very similar binding mode to the crystal structure is adopted (Figure 4.6 C and D, RMSDs of 1.1 Å and 1.5 Å respectively). More atoms in the ligand achieve favourable scores, as many more of the atoms are coloured pink. In particular, parts of the aromatic rings contribute to the score, indicating that the aromatic grid map has contributed to this improvement in scoring. This is confirmed by investigating the aromatic grid map as shown in Figure 4.7. The fused aromatic ring can be seen to be central to a large negative area of aromatic affinity.




Grid 3, N.C. 1, Aromatic (-3.5) Grid 3, N.C. 10, Aromatic (-25) Grid 3, N.C. 100, Carbon (-100)

Figure 4.7. AutoCRANkS grid maps for HSP90A using grid 3 and normalisation constants of 1 (A), 10 (B) and 100 (C). The isocontour level in kcal/mol is shown in brackets for each map. The docked pose for ligand PU3 from PDB structure 1uy6 is also shown. Aromatic grid maps are shown in A and B by the green mesh and an aliphatic carbon map is shown in C by the grey mesh.

However, as the AutoCRANkS normalisation constant increases further the RMSD worsens, to give docking pose that is further from the crystallographic structure than the pose generated by AutoDock. The conformer generated using a normalisation constant of 10 has now swapped the aromatic rings so that the benzyl ring is inside the well of affinity of the aromatic grid map. The aromatic fused ring system is now making the greatest contribution to the score and overlaps with the attractive well in the aromatic potential (Figure 4.7 B). The conformer generated using a normalisation constant of 100 is similar except an aliphatic branch has changed direction. The aliphatic carbon grid (Figure 4.7 C) shows why the tail has been twisted, as it now sits in a well of attractive aliphatic carbon affinity, worsening the accuracy of the predicted binding mode. This indicates that weighting the contribution of the active grids too strongly can be detrimental to successful binding mode prediction at these high normalisation constants.

#### *4.3.1.1.1 Re-Docking Summary*

To determine the effect of incorporating chemical and structural knowledge about active ligands into AutoGrid maps, to yield AutoCRANkS, ligands from PDB structures were redocked for each target. The results were assessed by investigating the RMSD between docked poses and the X-ray crystallography observed binding mode of the cognate ligand. Overall, results are promising. For targets with a large number of PDB structures to be redocked, there is an improvement in the average RMSD over conformers and in particular the RMSD between the structure and the lowest scored conformer. For example, for target HSP90A using a normalisation constant of 0.1 the proportion of ligands redocked with a crystallographic RMSD of less than 2 Å increases to over 50% when using AutoCRANkS compared to approximately 30% when using AutoDock. However, there is no consensus between targets at this stage for the optimal normalisation constant.

### *4.3.1.2 Docking of Actives and Decoys*

To investigate the performance of AutoCRANkS on the discrimination between actives and decoys, each of the selected DEKOIS 2.0 datasets were scored and ranked. Five different grids were used and added to AutoGrid maps using five different normalisation constants. The performance was tested using three metrics: Area Under the Receiver Operation Characteristic Curve (AUC), Enrichment Factor at 0.5% ($EF_{0.5\%}$), and BEDROC ($\alpha = 80.5$) (these metrics are detailed in Chapter 3, Section 3.2.3.

Figure 4.8 shows the calculated AUC for each target, including the average, best and worst result across the five grids for each normalisation constant. For targets HSP90A, ADRB2, DHFR and KIF11, on average there is an increase in AUC with a normalisation constant of 1 achieving an average AUC of 0.94 for KIF11. For DHFR, AutoDock with standard AutoGrid maps performs close to random (an AUC of 0.47), but when using AutoCRANkS and a normalisation constant of 1, an AUC of 0.78 is achieved – a significant improvement. For ADRB2, there is a slight increase in performance (an AUC of 0.65 when using AutoDock and an AUC of 0.74 when using AutoCRANkS with a normalisation constant of 1). For HSP90A, AutoDock achieves a worse than random ranking of actives with respect to decoys (AUC = 0.34). There is an increase in performance when using AutoCRANkS maps, on average using a normalisation constant of 1 an AUC of 0.57 is achieved – an increase from worse than random performance when using the original AutoDock protocol to better than random when using AutoCRANkS.

In contrast for BCL2, there is a large decrease in performance from 0.85 when using AutoDock with AutoGrid maps to worse than random when using AutoCRANkS, with a normalisation constant of 1 or higher (on average an AUC of between 0.42 to 0.43 was achieved for normalisation constants of 1 or higher). For AKT1, there is a slight decrease in performance, again at higher normalisation constants - when using AutoDock an AUC of 0.67 is achieved whereas when using AutoCRANkS on average of between 0.60 and 0.62 were achieved with a normalisation constant of 1 or higher.

| Target | Normalisation Constant | HSP90A | ADRB2 | DHFR | KIF11 | BCL2 | AKT1 |
|---|---|---|---|---|---|---|---|
| **AutoDock** | **0** | 0.31 | 0.65 | 0.47 | 0.72 | 0.85 | 0.67 |
| **Median** | **0.01** | 0.36 | 0.66 | 0.44 | 0.74 | **0.85** | 0.62 |
| | **0.1** | 0.47 | **0.74** | 0.69 | 0.88 | 0.76 | **0.65** |
| | **1** | 0.57 | 0.68 | 0.78 | **0.94** | 0.42 | 0.62 |
| | **10** | 0.53 | 0.71 | **0.79** | 0.93 | 0.44 | 0.60 |
| | **100** | **0.59** | 0.68 | 0.79 | 0.88 | 0.43 | 0.62 |
| **Best** | **0.01** | 0.36 | 0.69 | 0.45 | 0.76 | **0.86** | 0.63 |
| | **0.1** | 0.50 | **0.77** | 0.70 | 0.89 | 0.79 | **0.67** |
| | **1** | 0.71 | 0.73 | 0.81 | **0.97** | 0.44 | 0.65 |
| | **10** | 0.73 | 0.76 | **0.82** | 0.96 | 0.46 | 0.62 |
| | **100** | **0.75** | **0.77** | 0.81 | 0.93 | 0.46 | 0.63 |
| **Worst** | **0.01** | 0.34 | 0.64 | 0.43 | 0.72 | **0.84** | 0.59 |
| | **0.1** | 0.36 | **0.72** | 0.63 | 0.85 | 0.68 | **0.62** |
| | **1** | **0.48** | 0.60 | **0.74** | **0.91** | 0.41 | 0.55 |
| | **10** | 0.40 | 0.61 | 0.73 | 0.88 | 0.41 | 0.50 |
| | **100** | 0.39 | 0.62 | 0.72 | 0.87 | 0.41 | 0.50 |
| **Grid 1** | **0.01** | 0.34 | 0.69 | 0.44 | 0.72 | **0.84** | 0.63 |
| | **0.1** | 0.40 | **0.73** | 0.70 | 0.88 | 0.68 | **0.64** |
| | **1** | **0.48** | 0.63 | 0.81 | **0.94** | 0.42 | 0.62 |
| | **10** | 0.47 | 0.64 | **0.82** | **0.94** | 0.41 | 0.56 |
| | **100** | 0.47 | 0.65 | 0.81 | 0.90 | 0.41 | 0.50 |
| **Grid 2** | **0.01** | 0.36 | 0.69 | 0.44 | 0.72 | **0.85** | 0.63 |
| | **0.1** | 0.36 | **0.74** | 0.68 | 0.85 | 0.78 | **0.65** |
| | **1** | **0.48** | 0.68 | 0.77 | **0.91** | 0.41 | 0.63 |
| | **10** | 0.40 | 0.71 | **0.78** | 0.88 | 0.45 | 0.60 |
| | **100** | 0.39 | 0.68 | **0.78** | 0.87 | 0.46 | 0.63 |
| **Grid 3** | **0.01** | 0.35 | 0.66 | 0.44 | 0.74 | **0.86** | 0.59 |
| | **0.1** | 0.47 | 0.75 | 0.70 | 0.87 | 0.79 | **0.62** |
| | **1** | 0.57 | 0.73 | 0.78 | **0.95** | 0.43 | 0.55 |
| | **10** | 0.53 | 0.76 | **0.79** | 0.93 | 0.46 | 0.50 |
| | **100** | **0.59** | **0.77** | **0.79** | 0.87 | 0.43 | 0.51 |
| **Grid 4** | **0.01** | 0.36 | 0.65 | 0.43 | 0.76 | **0.85** | 0.62 |
| | **0.1** | 0.48 | **0.72** | 0.63 | 0.88 | 0.76 | **0.67** |
| | **1** | 0.71 | 0.60 | **0.74** | **0.94** | 0.44 | 0.65 |
| | **10** | **0.73** | 0.61 | 0.73 | 0.91 | 0.44 | 0.62 |
| | **100** | 0.75 | 0.62 | 0.72 | 0.88 | 0.43 | 0.62 |
| **Grid 5** | **0.01** | 0.36 | 0.64 | 0.45 | 0.74 | **0.84** | 0.61 |
| | **0.1** | 0.50 | **0.77** | 0.69 | 0.89 | 0.74 | **0.66** |
| | **1** | **0.61** | 0.71 | 0.80 | **0.97** | 0.42 | 0.62 |
| | **10** | **0.61** | 0.76 | **0.81** | 0.96 | 0.42 | 0.61 |
| | **100** | 0.59 | **0.77** | 0.80 | 0.93 | 0.41 | 0.62 |

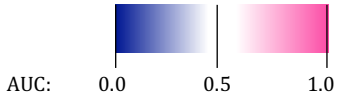



Figure 4.8. AUC for the DEKOIS actives and decoys sets for six targets using AutoCRANkS with five different normalisation constants. Five different grids protein-ligand structures were used. The average, best and worst results are shown across the five grids for each of the normalisation constants. When using a normalisation constant of 1 there is an improvement in AUC when using AutoCRANkS over AutoDock in four out of the six targets.

The $EF_{0.5\%}$ for the ranking of actives and decoys for each target is shown in Figure 4.9. The performance of AutoCRANkS is compared to AutoDock but also to AutoDock Vina, GOLD and Glide using values taken from the literature (Bauer *et al.,* 2013). On average for a normalisation constant of 1, AutoCRANkS outperforms all four docking tools significantly for ADRB2 ($EF_{0.5\%}$ = 21 compared to $EF_{0.5\%}$ = 5 for AutoDock Vina, GOLD and GLIDE), DHFR ($EF_{0.5\%}$ = 10, compared to $EF_{0.5\%}$ = 0 for AutoDock Vina and GOLD, and $EF_{0.5\%}$ = 5 for Glide) and KIF11 ($EF_{0.5\%}$ = 31 compared to $EF_{0.5\%}$ = 21 for AutoDock Vina, $EF_{0.5\%}$ = 15 for GOLD. and $EF_{0.5\%}$ = 5 for Glide).

However, for HSP90A, BCL2 and AKT1, on average zero enrichment was found for all normalisation constants > 0.01 when using AutoCRANkS. For HSP90A, only GOLD achieved enrichment above zero and this was only an enrichment of 5. For AKT1, although AutoDock Vina, Glide and GOLD achieved high enrichment (ranging from 5 to 20), AutoDock also achieved an enrichment of 0. BCL2, in contrast, was found to have a high enrichment across all other docking tools ranging from 5 to 20.

| Target | | HSP90A | ADRB2 | DHFR | KIF11 | BCL2 | AKT1 |
|---|---|---|---|---|---|---|---|
| | AutoDock Vina | 0 | **5** | 0 | **21** | **15** | 11 |
| | Gold | **5** | **5** | 0 | 15 | 5 | 5 |
| | Glide | 0 | **5** | **5** | 5 | 5 | **20** |
| | Normalisation Constant | | | | | | |
| AutoDock | **0** | 0 | 5 | 0 | 10 | 20 | 0 |
| Median | **0.01** | 0 | 5 | 0 | 6 | **11** | 0 |
| | **0.1** | 0 | 5 | 5 | 21 | 0 | 0 |
| | **1** | 0 | **21** | 10 | **31** | 0 | 0 |
| | **10** | 0 | 11 | **15** | 26 | 0 | 0 |
| | **100** | 0 | 6 | **15** | 5 | 0 | 0 |
| Best | **0.01** | 0 | 10 | 0 | 10 | **25** | 0 |
| | **0.1** | 0 | 15 | 6 | 26 | 5 | 5 |
| | **1** | 5 | **25** | **20** | **31** | 5 | **10** |
| | **10** | 5 | 20 | 16 | **31** | 1 | 5 |
| | **100** | **11** | 15 | 15 | **31** | 0 | 1 |
| Worst | **0.01** | 0 | 5 | 0 | 0 | **6** | 0 |
| | **0.1** | 0 | 5 | 5 | **15** | 0 | 0 |
| | **1** | 0 | **10** | **10** | 5 | 0 | 0 |
| | **10** | 0 | **10** | 6 | 5 | 0 | 0 |
| | **100** | 0 | 5 | 5 | 0 | 0 | 0 |
| Grid 1 | **0.01** | 0 | 5 | 0 | 5 | **11** | 0 |
| | **0.1** | 0 | 5 | 6 | 20 | 0 | 0 |
| | **1** | 0 | **25** | **20** | **31** | 0 | **10** |
| | **10** | 0 | 10 | 16 | 30 | 0 | 0 |
| | **100** | 0 | 6 | 15 | 15 | 0 | 0 |
| Grid 2 | **0.01** | 0 | 5 | 0 | 6 | **11** | 0 |
| | **0.1** | 0 | 5 | 5 | **15** | 5 | 0 |
| | **1** | 0 | **21** | 11 | 5 | 0 | 0 |
| | **10** | 0 | 20 | **15** | 5 | 1 | 0 |
| | **100** | 0 | 6 | **15** | 0 | 0 | 0 |
| Grid 3 | **0.01** | 0 | 10 | 0 | 10 | **25** | 0 |
| | **0.1** | 0 | 5 | 5 | **21** | 5 | 0 |
| | **1** | 5 | **21** | 10 | 10 | 0 | 0 |
| | **10** | 0 | 10 | **16** | 10 | 0 | 0 |
| | **100** | 0 | 5 | 15 | 5 | 0 | 0 |
| Grid 4 | **0.01** | 0 | 8 | 0 | 0 | **16** | 0 |
| | **0.1** | 0 | 15 | 5 | 26 | 0 | 0 |
| | **1** | 1 | **20** | **10** | **31** | 0 | **5** |
| | **10** | **5** | **20** | **10** | **31** | 0 | **5** |
| | **100** | **11** | 15 | 5 | **31** | 0 | 1 |
| Grid 5 | **0.01** | 0 | 5 | 0 | 10 | **6** | 0 |
| | **0.1** | 0 | **15** | 5 | 21 | 0 | **5** |
| | **1** | 0 | 10 | **10** | **31** | 5 | 0 |
| | **10** | 0 | 11 | 6 | 26 | 0 | 0 |
| | **100** | 0 | 5 | 5 | 0 | 0 | 0 |


Figure 4.9. $EF_{0.5\%}$ for the DEKOIS actives and decoys sets for six targets using AutoCRANkS with five different normalisation constants. Five different grids of protein-ligand structures were used. The average, best and worst results are shown across the five grids for each normalisation constant. Results for AutoDock Vina, GOLD and Glide taken from the literature are also shown (Bauer *et al.*, 2013). Using a normalisation constant of 1 a higher enrichment than all other docking methods compared to here was achieved for half of the targets.

Figure 4.10 shows the performance of AutoCRANkS with respect to AutoDock using BEDROC (α = 80.5). A similar behaviour is observed. For a normalisation constant of 1, AutoCRANkS performs significantly better than AutoDock with AutoGrid maps for ADRB2, DHFR and KIF11. For ADRB2 the BEDROC (α = 80.5) increases from 0.14 when using AutoDock to 0.40 on average across the grids when using AutoCRANkS with normalisation constant 1, for DHFR from 0.02 to 0.27 and for KIF11 from 0.20 to 0.79. For AKT1, there is a slight improvement from BEDROC (α = 80.5) = 0.01 when using AutoDock to 0.09 compared to using AutoCRANkS with a normalisation constant of 1. For HSP90A the BEDROC remains at zero on average. In contrast, for BCL2, AutoCRANkS performs significantly worse than AutoDock with a BEDROC of zero for a normalisation constant of 1, whilst AutoDock achieves a BEDROC of 0.48.

In summary, AutoCRANkS generally outperformed AutoGrid maps with AutoDock as measured by AUC, $EF_{0.5\%}$ and BEDROC (α = 80.5). A normalisation constant of 1 was found to be on average most successful. However, the performance of AutoCRANkS was found to be highly target-dependent. For targets for which AutoCRANkS performed extremely well (three out of the six targets: ADRB2, DHFR and KIF11), it also outperformed AutoDock, AutoDock Vina, GOLD and Glide when using a normalisation constant of 1. However, for one target there was a slight decrease in performance for AUC and $EF_{0.5\%}$ (AKT1) and for one target a significant decrease in performance (BCL2). For target HSP90A there was an increase in the AUC when using AutoCRANkS with a normalisation constant compared to AutoDock but the enrichment and BEDROC (α = 80.5) remained at zero. The performance is explored further for a selection of targets below.

| Target | Normalisation Constant | HSP90A | ADRB2 | DHFR | KIF11 | BCL2 | AKT1 |
|---|---|---|---|---|---|---|---|
| **AutoDock** | **0** | 0.00 | 0.14 | 0.02 | 0.20 | 0.48 | 0.01 |
| **Median** | **0.01** | 0.00 | 0.17 | 0.03 | 0.19 | **0.46** | 0.01 |
| | **0.1** | 0.00 | 0.17 | 0.09 | 0.46 | 0.04 | 0.04 |
| | **1** | 0.00 | **0.40** | 0.27 | **0.79** | 0.00 | **0.09** |
| | **10** | **0.02** | 0.27 | **0.33** | 0.58 | 0.01 | 0.07 |
| | **100** | 0.00 | 0.21 | 0.29 | 0.25 | 0.00 | 0.01 |
| **Best** | **0.01** | 0.00 | 0.21 | 0.03 | 0.22 | **0.59** | 0.04 |
| | **0.1** | 0.00 | 0.31 | 0.17 | 0.69 | 0.20 | 0.08 |
| | **1** | 0.01 | **0.47** | **0.40** | **0.86** | 0.07 | **0.20** |
| | **10** | **0.12** | 0.37 | 0.39 | 0.84 | 0.05 | 0.10 |
| | **100** | 0.09 | 0.25 | 0.39 | 0.80 | 0.04 | 0.09 |
| **Worst** | **0.01** | 0.00 | 0.11 | 0.00 | 0.18 | **0.39** | 0.00 |
| | **0.1** | 0.00 | 0.12 | 0.07 | **0.40** | 0.00 | **0.02** |
| | **1** | 0.00 | **0.30** | 0.16 | 0.20 | 0.00 | 0.01 |
| | **10** | 0.00 | 0.25 | **0.18** | 0.15 | 0.00 | 0.00 |
| | **100** | 0.00 | 0.15 | 0.16 | 0.11 | 0.00 | 0.00 |
| **Grid 1** | **0.01** | 0.00 | 0.11 | 0.03 | 0.22 | **0.39** | 0.04 |
| | **0.1** | 0.00 | 0.15 | 0.17 | 0.55 | 0.00 | 0.03 |
| | **1** | 0.00 | **0.41** | **0.40** | **0.86** | 0.00 | **0.20** |
| | **10** | 0.00 | 0.26 | 0.39 | 0.81 | 0.00 | 0.07 |
| | **100** | 0.00 | 0.15 | 0.39 | 0.50 | 0.00 | 0.00 |
| **Grid 2** | **0.01** | 0.00 | 0.21 | 0.03 | 0.21 | **0.49** | 0.03 |
| | **0.1** | 0.00 | 0.12 | 0.08 | **0.40** | 0.17 | 0.04 |
| | **1** | 0.00 | **0.47** | 0.28 | 0.20 | 0.00 | **0.09** |
| | **10** | 0.02 | 0.37 | **0.33** | 0.15 | 0.05 | 0.07 |
| | **100** | 0.00 | 0.25 | 0.30 | 0.11 | 0.01 | 0.02 |
| **Grid 3** | **0.01** | 0.00 | 0.15 | 0.00 | 0.19 | **0.59** | 0.01 |
| | **0.1** | 0.00 | 0.17 | 0.07 | **0.44** | 0.20 | **0.02** |
| | **1** | 0.01 | **0.40** | 0.23 | 0.39 | 0.00 | 0.01 |
| | **10** | **0.10** | 0.25 | **0.33** | 0.29 | 0.02 | 0.00 |
| | **100** | 0.08 | 0.17 | 0.29 | 0.17 | 0.00 | 0.00 |
| **Grid 4** | **0.01** | 0.00 | 0.19 | 0.03 | 0.18 | **0.46** | 0.00 |
| | **0.1** | 0.00 | 0.27 | 0.09 | 0.69 | 0.02 | 0.04 |
| | **1** | 0.01 | **0.34** | 0.16 | **0.85** | 0.02 | **0.10** |
| | **10** | **0.12** | 0.27 | **0.18** | 0.84 | 0.01 | **0.10** |
| | **100** | 0.09 | 0.24 | **0.18** | 0.80 | 0.04 | 0.09 |
| **Grid 5** | **0.01** | 0.00 | 0.17 | 0.03 | 0.19 | **0.43** | 0.01 |
| | **0.1** | 0.00 | **0.31** | 0.12 | 0.46 | 0.04 | **0.08** |
| | **1** | 0.00 | 0.30 | **0.27** | **0.79** | 0.07 | 0.04 |
| | **10** | **0.02** | 0.27 | 0.23 | 0.58 | 0.01 | 0.02 |
| | **100** | 0.00 | 0.21 | 0.16 | 0.25 | 0.00 | 0.01 |

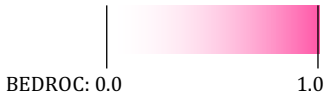



Figure 4.10. BEDROC (α= 80.5) for the DEKOIS actives and decoys sets for six targets using AutoCRANkS with five different normalisation constants. Five different grids protein-ligand structures were used. The average, best and worst results are shown across the five grids for each of the normalisation constants. For ADRB2, KIF11 and DHFR for a normalisation constant of 1 AutoCRANkS achieves higher BEDROC (α= 80.5) than when using AutoDock.

#### *4.3.1.2.1 HSP90A*

HSP90A (heat shock protein 90-alpha) is a chaperone protein targeted due its function stabilising a number of proteins necessary for the growth of tumours. There are four water molecules found to be conserved in the binding site that are important to consider when designing inhibitors for this target (Yan *et al.,* 2008). These are shown in Figure 4.11. Additionally, there is a flexible helix that interacts with a number of waters and inhibitors of the target, shown in pink in Figure 4.11. This can make it a difficult target to dock to.

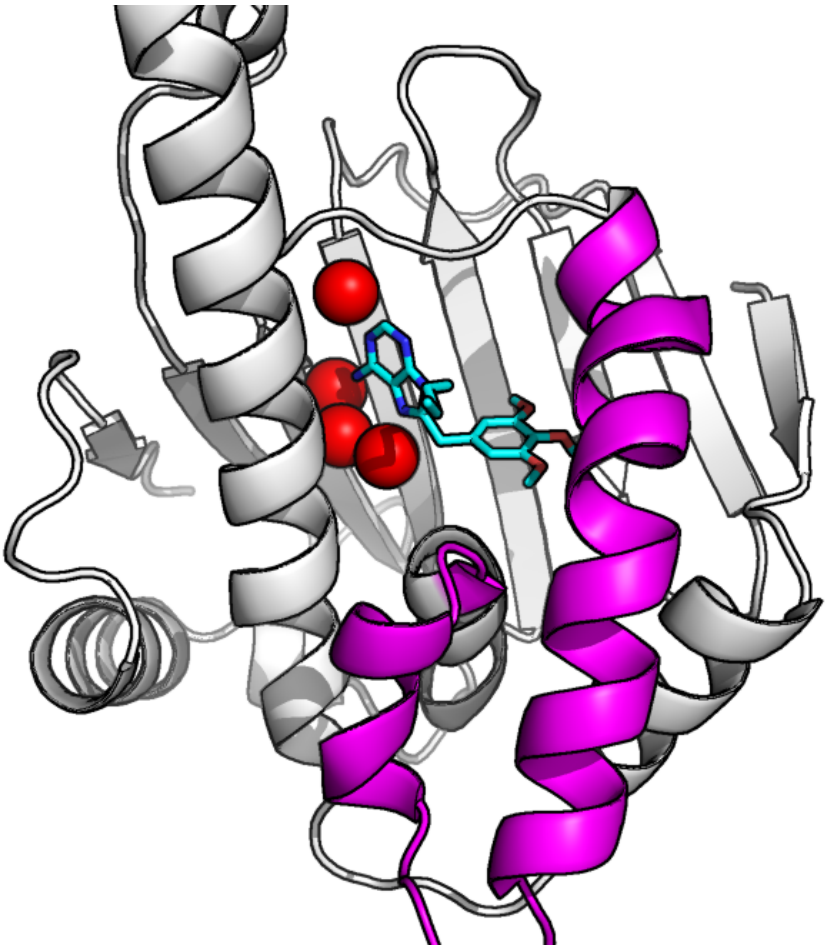


Figure 4.11. PDB structure 1uy6 for target HSP90A. Structurally conserved waters important for binding are shown by red spheres and the flexible helix is coloured pink.

Figure 4.12 shows the structure used in the docking protocol with the waters that were included in the calculation. The waters within the crystal structure found within 2 Å of the waters docked by WaterDock were kept in the complex for docking and all others were removed. This method has worked well with three out of four of the waters known to be important for binding remain in the structure. Only one of the conserved water molecules was removed.

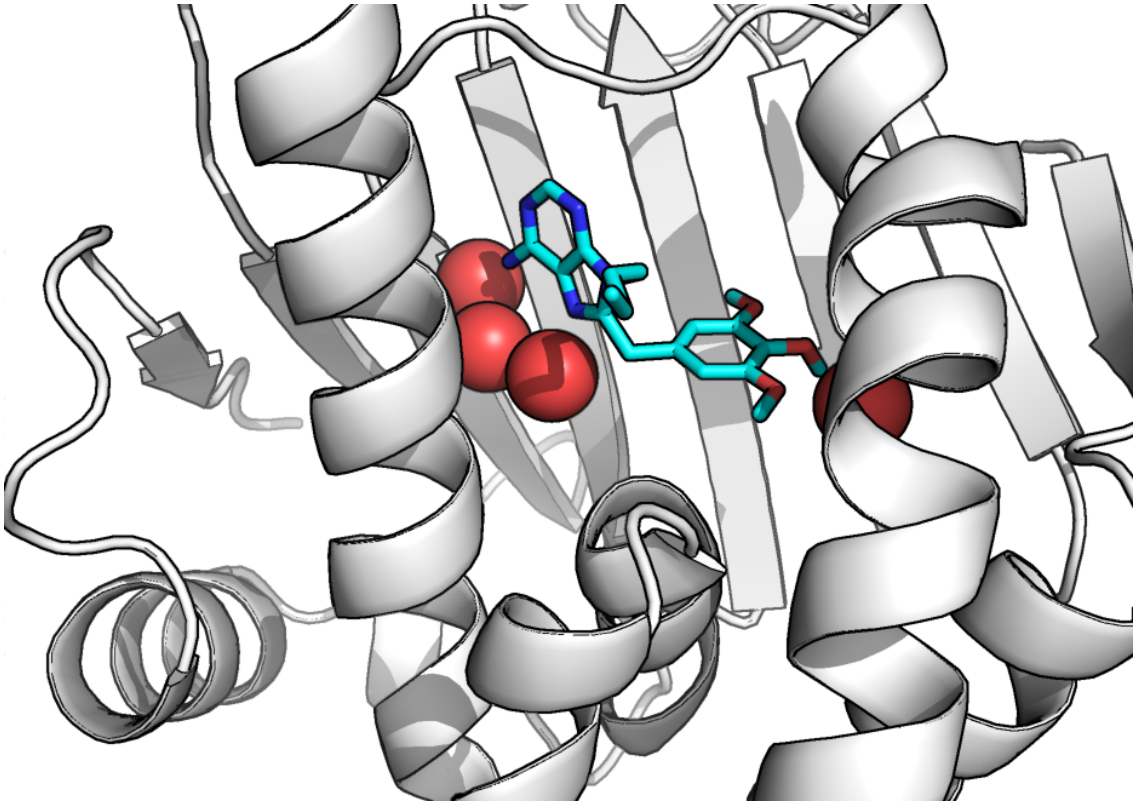


Figure 4.12. PDB structure 1uy6 for target HSP90A after the pre-docking protocol. Waters calculated to remain in the structure are shown by red spheres. Three out of four of the structurally conserved waters shown in Figure 4.11 are kept within the structure. The ligand PU3 is also shown although this is not included in the structure used for docking.

For this target, there was a large improvement in performance on using AutoCRANkS over AutoDock as measured by AUC, particularly for a normalisation constant of 1 (on average an AUC of 0.57 compared to an AUC of 0.31 for AutoDock). However, for $EF_{0.5\%}$ there is no improvement with zero enrichment achieved by both AutoDock and AutoCRANkS and for BEDROC ($\alpha = 80.5$) there is little improvement.

The performance is grid dependent with the best results achieved for grid 3 and the worst for grid 1. To explore this, Figures 4.13 and 4.14 show histograms of the scores for each run for HSP90A using grid 1 and grid 3 respectively. A marked difference in behaviour can be seen between the two grids. For grid 1, there is a little variation across the distributions as the normalisation constant increases. The shape is relatively similar apart from a tail to the right of the distribution which increases as the normalisation constant increases.

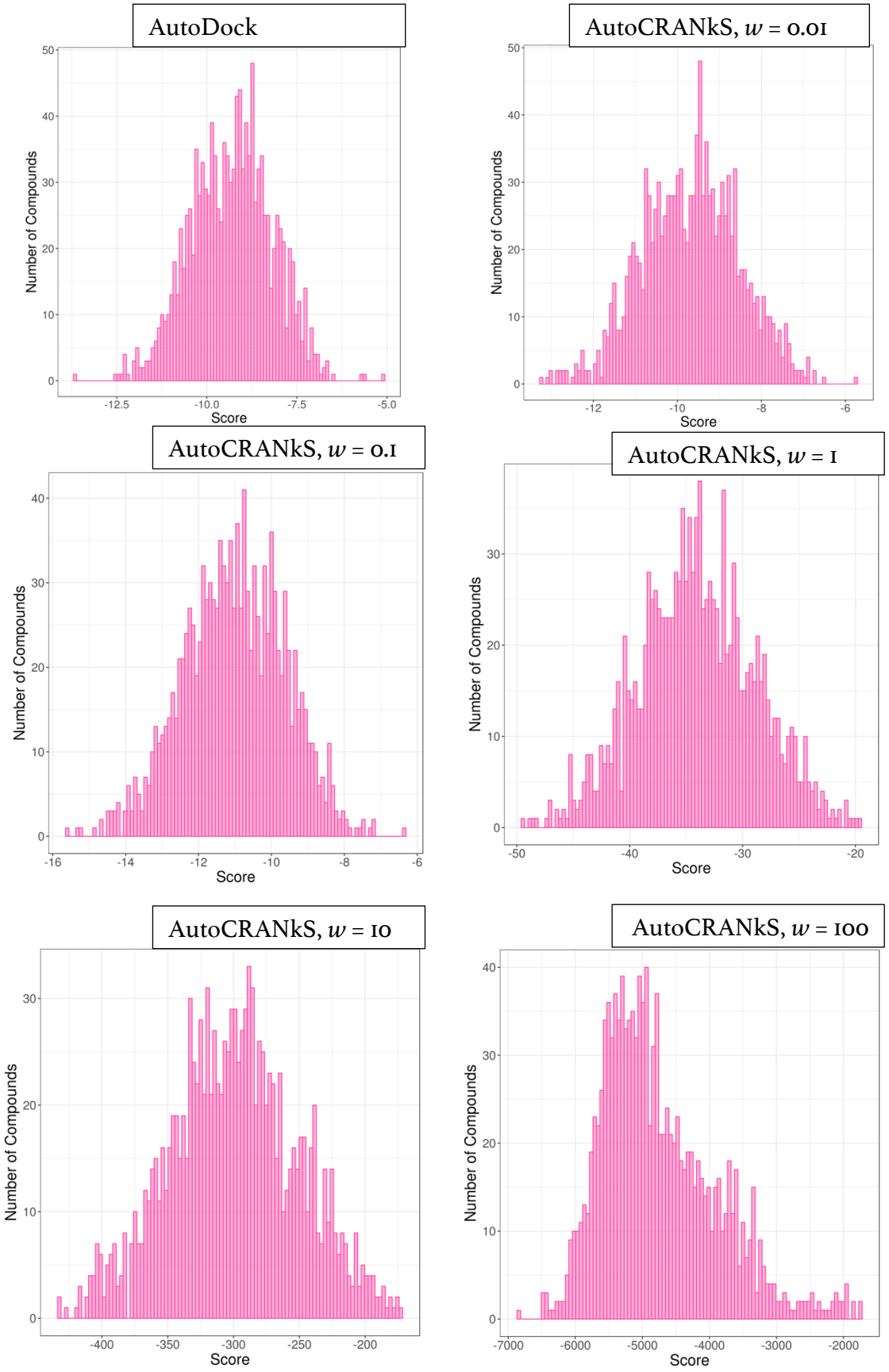



Figure 4.13. Distributions of scores for docking of actives and decoys from HSP90A DEKOIS dataset. Compounds were docked using AutoDock and then AutoCRANkS using grid 1 and five normalisation constants.

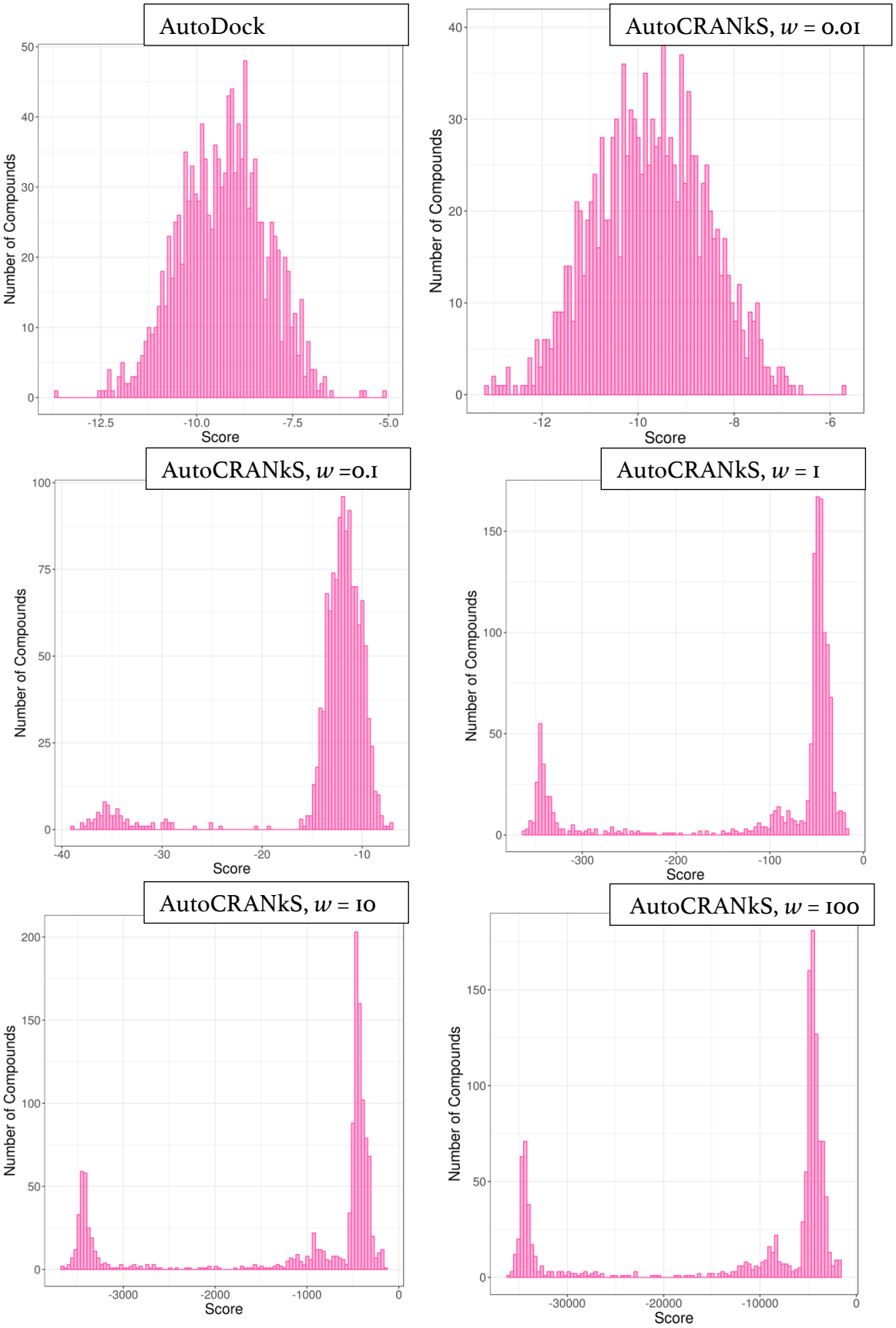



Figure 4.14. Distributions of scores for docking of actives and decoys from HSP90A DEKOIS dataset. Compounds were docked using AutoDock and then AutoCRANkS using grid 3 and five normalisation constants.

However, the histograms in grid 3 are very different to each other across normalisation constants. As the normalisation constant increases, the distribution becomes bimodal with a normalisation constant of 1 and could be described as trimodal when the normalisation constant increases to 10 and 100. This could explain the increase in performance as the scores are split into two groups, with more decoys in the higher scoring group and more actives in the lower scoring group. For example, for a normalisation constant of 1 if a cut-off of a score of -300 is used this should include all molecules in the first peak. This set contains 206 compounds of which 6% are active. If molecules with a score of greater than -50 is used this should cover compounds in the peak on the right of Figure 4.14. In this set there were 635 compounds of which 3% are active. This exhibits the change in proportion of actives and decoys between the two peaks which yields the increased performance of AutoCRANkS. The fact there is zero enrichment calculated by $EF_{0.5\%}$ and BEDROC ($\alpha = 80.5$) indicates it is the down-ranking of decoys that causes the increase in performance rather than the up-ranking of actives.

To investigate this further, an example of a docked active (active compound 4 from the DEKOIS 2.0 dataset for HSP90A) by AutoCRANkS using grid 3 and a normalisation constant of 10, by AutoCRANkS using grid 1 and a normalisation of 10 and by AutoDock is shown in Figure 4.15. The compound is coloured by the atomic affinity score for each atom, allowing a visualisation of how each atom contributes to the score. The corresponding grid maps for each of the poses are also shown in Figure 4.16.

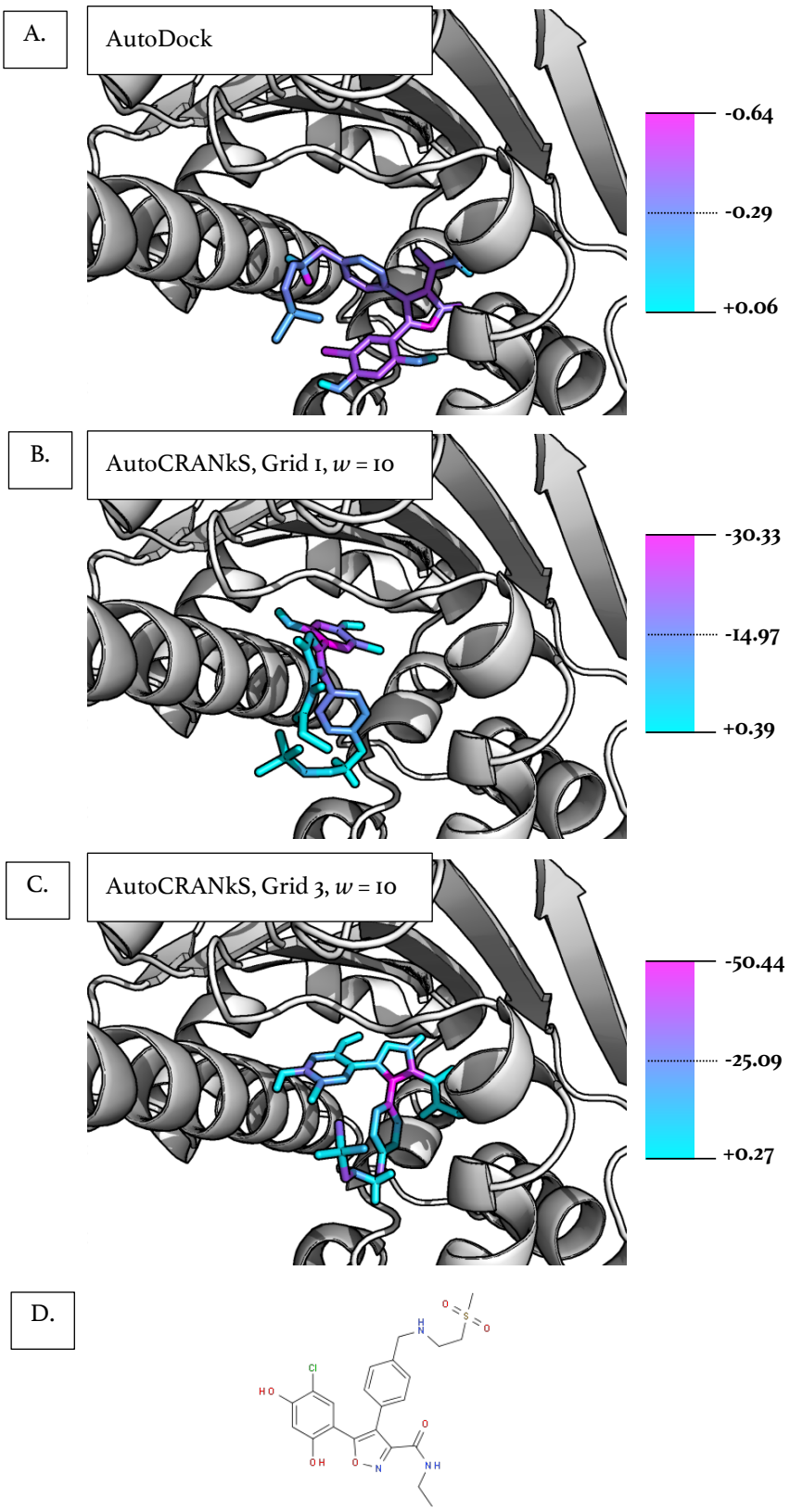



Figure 4.15. Docked active compound number 4 from the HSP90A DEKOIS dataset by AutoDock (A) and AutoCRANkS using grids 1 (B) and 3 (C) with a normalisation constant of 10. The docked poses are coloured by the per atom contribution to the score. The 2D structure is shown in D. The addition of the grids causes a different conformation and scoring pattern when using AutoCRANkS.

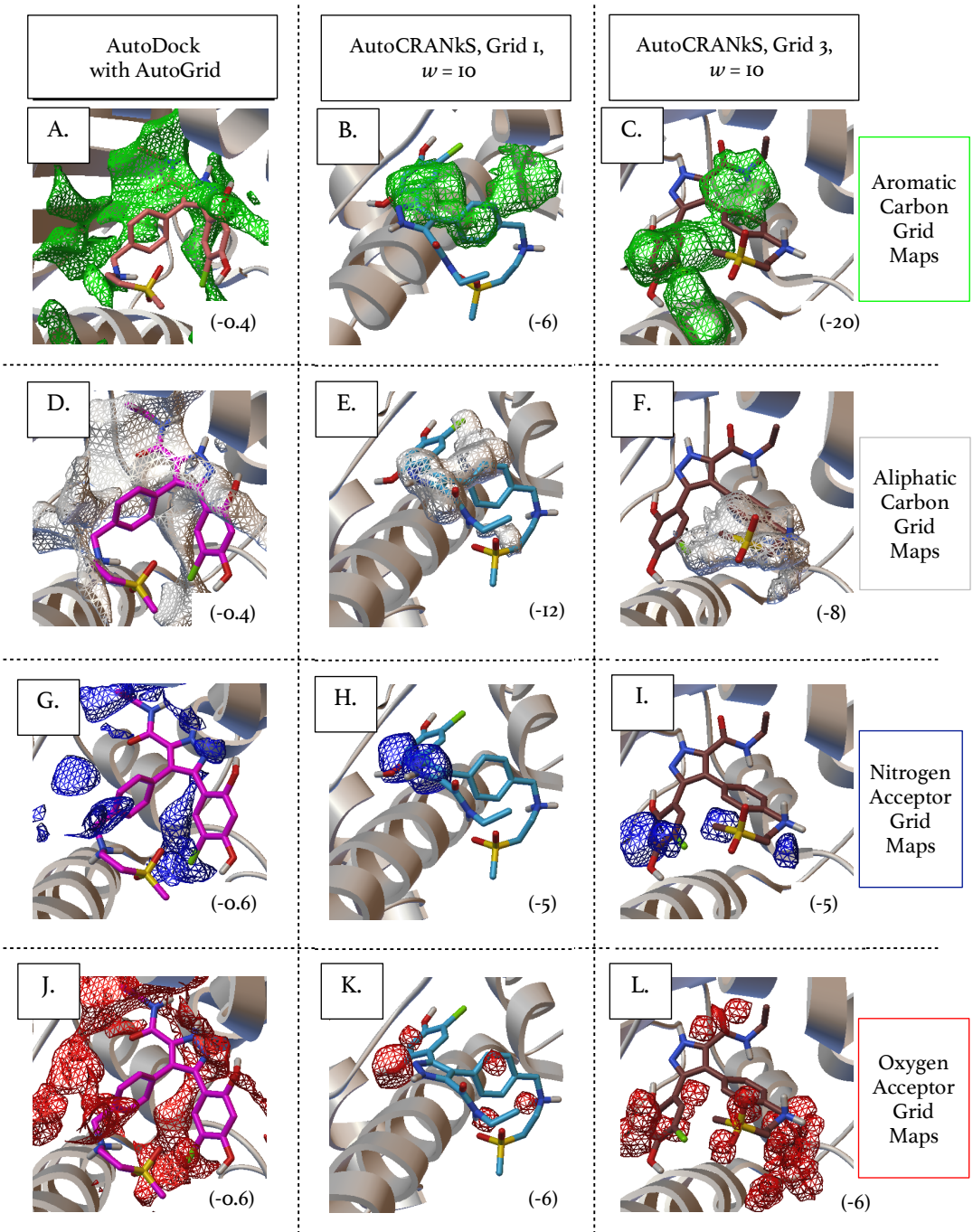



Figure 4.16. Aromatic carbon grid maps are shown with poses generated by AutoDock (A), AutoCRANkS grid 1, $w$ = 10 (B) and AutoCRANkS grid 3, $w$ = 10 (C). Aliphatic carbon grid maps are shown with poses generated by AutoDock (D), AutoCRANkS grid 1, $w$ = 10 (E) and AutoCRANkS grid 3, $w$ = 10 (F). Nitrogen acceptor grid maps are shown with poses generated by AutoDock (G), AutoCRANkS grid 1, $w$ = 10 (H) and AutoCRANkS grid 3, $w$ = 10 (I). Oxygen acceptor maps are shown with poses generated by AutoDock (J), AutoCRANkS grid 1, $w$ = 10 (K) and AutoCRANkS grid 3, $w$ = 10 (L). The addition of the CRANkS grids shows distinct changes to the grid maps. The higher number of attractive wells in the oxygen acceptor map (L) and aromatic map in grid 3 (C) cause the molecule to be ranked higher by AutoCRANkS using grid 3, $w$ = 10. The isocontour level (in kcal/mol) for each map is shown in brackets. The docked compound is active compound 4 from the DEKOIS 2.0 for HSP90A and 2D structure of the compound is shown in Figure 4.15 D.

All three poses are very different. For the AutoDock generated pose (Figure 4.15 A), a nitrogen on the pyrole ring makes the largest contribution to the score. The AutoGrid maps show the docked compound overlaps with a hotspot of attractive potential on the nitrogen acceptor map (Figure 4.16 G). The aromatic rings also show contribution to the favourable score and there is little overlap between the pyrole ring and attractive parts of the aromatic AutoGrid map (Figure 4.16 A).

The poses generated by the AutoCRANkS grids are also scored differently to each other and the AutoDock pose (Figure 4.15). For the pose generated by grid 1, the phenol-based ring contributes significantly to the score, as does part of the pyrole ring, including the nitrogen (Figure 4.15 B). The grid shows the overlap of the phenol ring with the aromatic grid map (Figure 4.16 B). There is also overlap between the nitrogen of the pyrole ring with attractive sections of the nitrogen acceptor map (Figure 4.16 H).

For the pose generated by AutoCRANkS using grid 3, the most significant contribution to the score is made by the linker between the pyrazole ring and one of the benzene rings (Figure 4.15 C). The overlap between this linker and a well of attractive potential in the aromatic map is shown to be the cause of this (Figure 4.16 C). A carbon atom linking the benzene ring with the amine group is also coloured pink and thus makes a large contribution to the score (Figure 4.15 C). The atom clearly overlaps with the aliphatic carbon grid map (Figure 4.16 F). There is also overlap with oxygen atoms and the oxygen acceptor grid map which contributes to the low score (Figure 4.16 J). However, there is no overlap with the nitrogen acceptor map (Figure 4.16 G).

This active compound (active compound 4 from the DEKOIS 2.0 dataset) was ranked 479$^{th}$ out of 1240 by AutoDock, 185$^{th}$ by AutoCRANkS using grid 1 with a normalisation constant of 1, and 36$^{th}$ AutoCRANkS using grid 3 with a normalisation constant of 1. The improvement is clearly caused by the change in actives used to compute the AutoCRANkS grid maps, which has aided the ranking. Between grid 1 and grid 3, the change in the aromatic map is very clear (Figure 4.16 B and C), as is the change in the oxygen acceptor map (Figure 4.16 K and L). The presence of favourable regions in the aromatic and oxygen acceptor AutoCRANkS maps in locations that overlap with the docked conformations of the ligand, as well as the carbon map, when using grid 3 could be the reason for the improved performance over using grid 1. This highlights the importance of the choosing the optimal set of bound active ligands to create the AutoCRANkS grids.

#### *4.3.1.2.2 DHFR*

Dihydrofolate Reductase (DHFR) is an enzyme that has been targeted by drugs as it is crucial for DNA synthesis and cell proliferation and so could be inhibited to combat tumours. For the actives and decoys set for DHFR, there is a large improvement in performance of discrimination between actives and decoys on using AutoCRANkS for normalisation constants of 1 or greater over AutoDock. In terms of AUC AutoDock achieved 0.47 whereas AutoCRANkS on average over the grids for a normalisation constant of 1 achieved an AUC of 0.78 (Figure 4.8). There is a similar performance across all five grids – for example for a normalisation constant of 1 the calculated AUC values ranged from 0.74 to 0.81 (Figure 4.8).

An example of a docked pose of an active compound (active compound 9 from the DEKOIS 2.0 DHFR dataset: 5-(5-Chloro-2,4-dihydroxyphenyl)-N-ethyl-4-[4-({[2-(methylsulfonyl)ethyl]amino}methyl)phenyl]-1,2-oxazole-3-carboxamide) generated by AutoDock (A) and by AutoCRANkS for grids 1 (B) and 4 (C) with a normalisation constant of 1 is shown in Figure 4.17. The corresponding grid maps are shown in Figure 4.18. This active compound is ranked 177$^{th}$ using AutoDock, 312$^{nd}$ when using AutoCRANkS with grid 4 and 4$^{th}$ when using AutoCRANkS with grid 1. Interestingly, despite similar performance for the grids in terms of AUC, $EF_{0.5\%}$ and BEDROC ($\alpha$ = 80.5) some actives can be scored well by some grids and not by others. Most actives have similar rankings by the two grids, and the slight increase in performance of grid 1 can be attributed to cases like this.

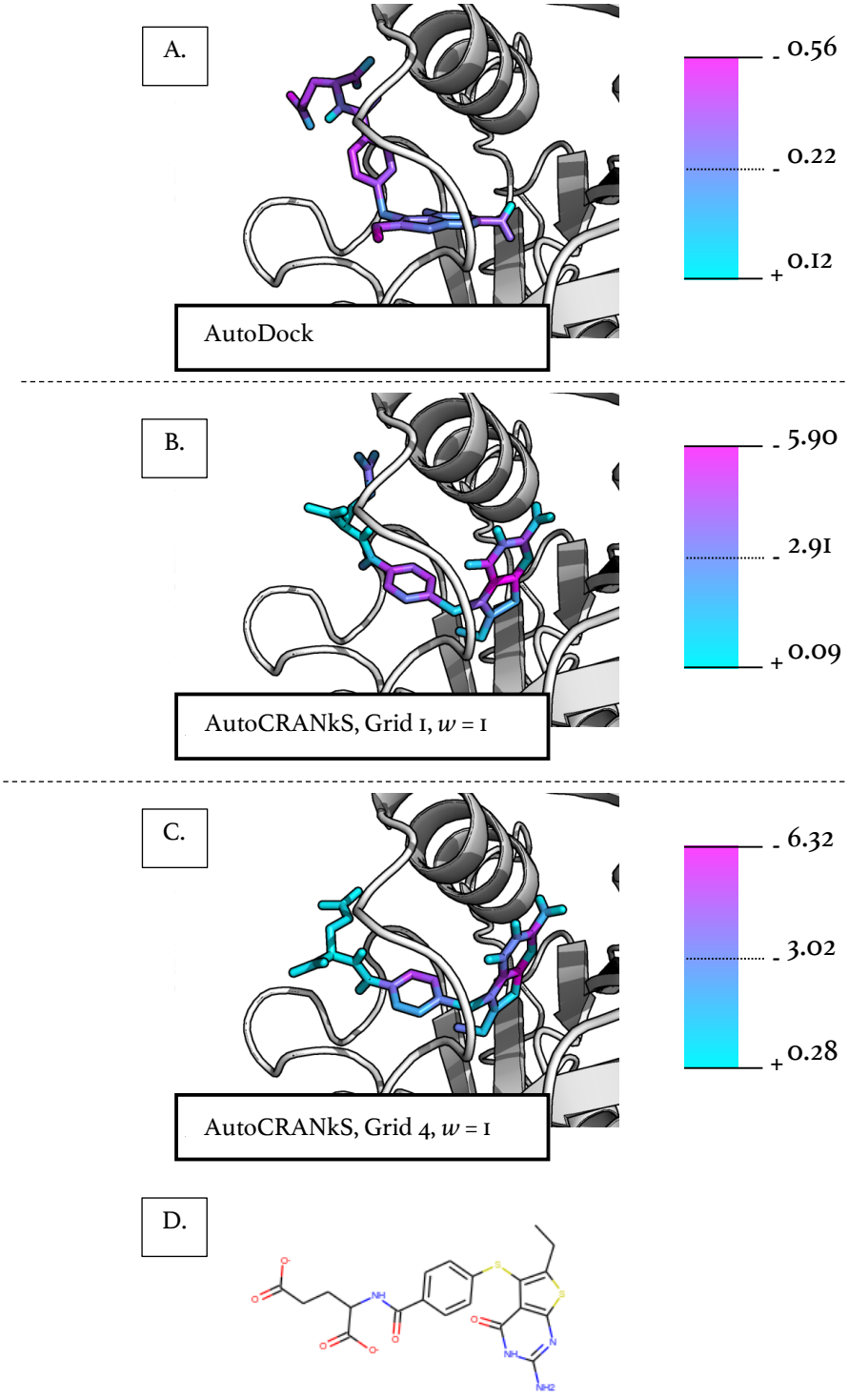



Figure 4.17. Docked poses for active compound 9 from the DHFR DEKOIS 2.0 dataset. Poses are shown for the lowest scored conformer docked by AutoDock (A) and AutoCRANkS using grids 1 (B) and 4 (C) with a normalisation constant of 1. The poses are coloured by the per-atom contribution to the score. The conformations generated by both AutoCRANkS grids are similar, but the scoring patterns are different which leads to different rankings of the compounds by the two grids.

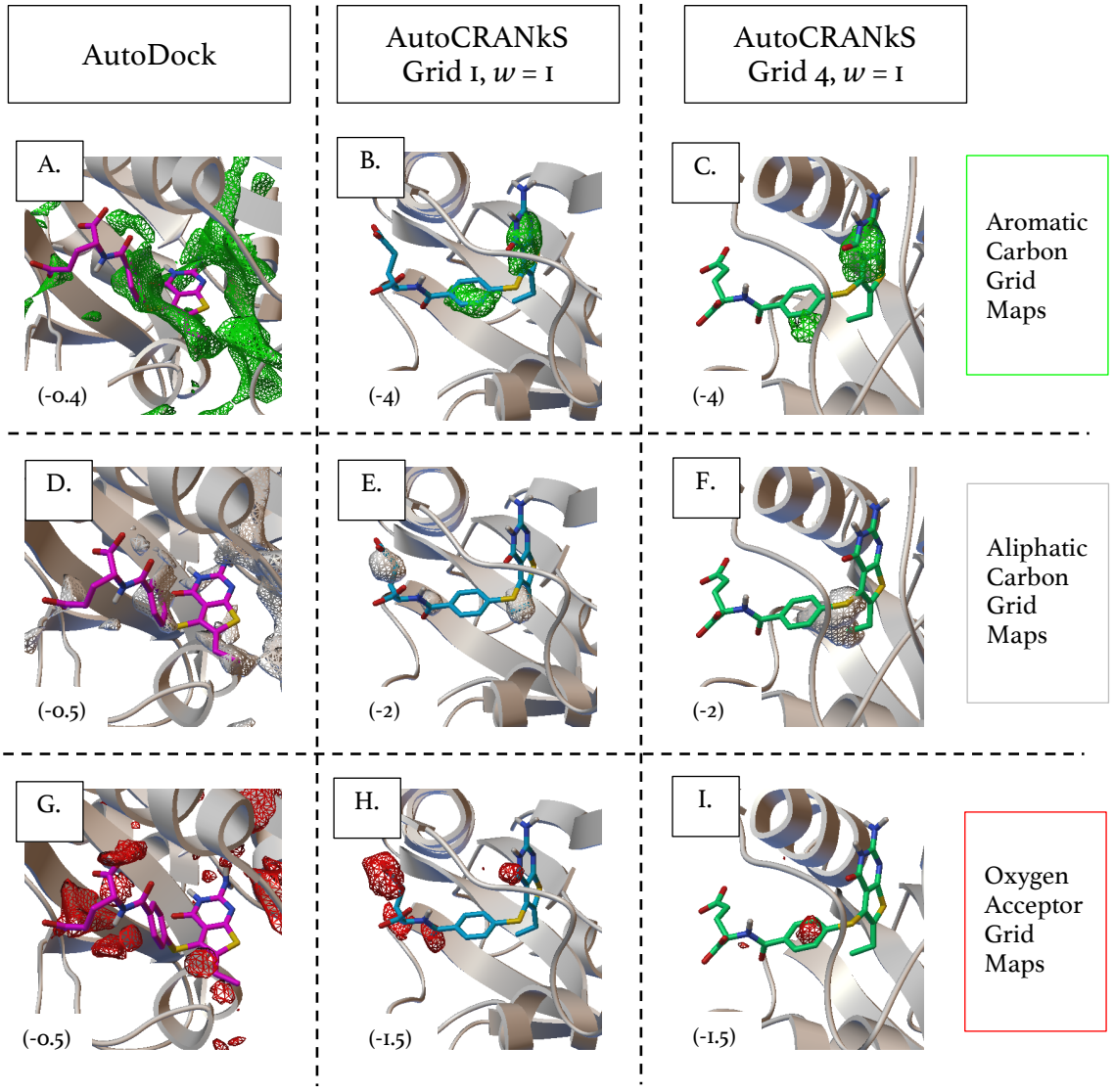



Figure 4.18. Docked poses for active compound 9 from the DEKOIS 2.0 DHFR dataset are shown to illustrate overlap with the grid maps. The isocontour level (in kcal/mol) for each map is shown in brackets. Aromatic carbon grid maps and corresponding docked poses are shown: used by AutoDock (A), AutoCRANkS grid 1, $w = 1$ (B) and AutoCRANkS grid 4, $w = 1$ (C). Aliphatic carbon grid maps and corresponding docked poses are shown: used by AutoDock (D), AutoCRANkS grid 1, $w = 1$ (E) and AutoCRANkS grid 4, $w = 1$ (F) . Oxygen acceptor grid maps and corresponding docked poses are shown: used by AutoDock (G), AutoCRANkS grid 1, $w = 1$ (H) and AutoCRANkS grid 4, $w = 1$ (I).

The docked poses for the two AutoCRANkS generated conformations are extremely similar – the only significant change is the branch finishing with alcohol groups (Figure 4.17). It is the scoring of the atoms that is affecting the ranking of the compound by the two methods. For grid 1, Figure 4.17 B shows that the score can be attributed to aromatic groups, and carbon and oxygen atoms on one of the diol

branches. The grid maps confirm this, as the two aromatic rings each overlap with an attractive well of the aromatic grid map (Figure 4.18 B). The carbon grid shows that the carboxylate branch is held in position by overlap with an area of attractive potential in the aliphatic carbon map (Figure 4.18 E) and this allows the oxygen atoms to also overlap with an attractive area of the oxygen acceptor map (Figure 4.18 H). In contrast for grid 4 there is only overlap with one aromatic attractive well, the other ring does not overlap (Figure 4.18 C). There is also no overlap with the carbon or oxygen acceptor grid maps (Figure 4.18 F and I), so the active compound is scored much higher for grid 4 than grid 1 and is not recognised as active when using grid 4.

#### *4.3.1.2.3 KIF11*

Kinesin-like protein 11 or KIF11 is a motor protein. It has been targeted by drugs due to its importance in the proliferation and invasion of glioblastoma (Venere *et al.,* 2015). High AUC values are achieved for KIF11 when using AutoCRANkS with a normalisation constant of 1. The AUC is increased from 0.72 for AutoDock to a maximum of 0.97 for grid 5. Active-guided docking thus improves the performance of the scoring of actives with respect to decoys for this target. However, as the normalisation constant increases above 1, the AUC begins to decline again as the active grids start to completely dominate the score (for grid 5 with a normalisation constant of 100 an AUC of 0.93 is achieved compared to 0.97 for a normalisation constant of 1).

An example of this can be seen in Figure 4.19, for a docked pose of active compound 28 from the DEKOIS 2.0 KIF11 dataset using AutoCRANkS with grid 3 and a normalisation constant of 1 (Figure 4.19 A) and a normalisation constant of 100 (Figure 4.19 B). The pose is coloured by the per atom contribution to the score. The corresponding grid maps are also shown in Figure 4.20. This active compound is ranked 11$^{th}$ when using a normalisation constant of 1 but 144$^{th}$ when using a normalisation constant of 100. For a normalisation constant of 100 as the active grids dominate, the tail of the molecule is forced round into a high energy orientation to match up with an attractive carbon well (Figure 4.20 D). This is likely to cause the higher score and consequently poor ranking of this molecule when using the higher normalisation constant.

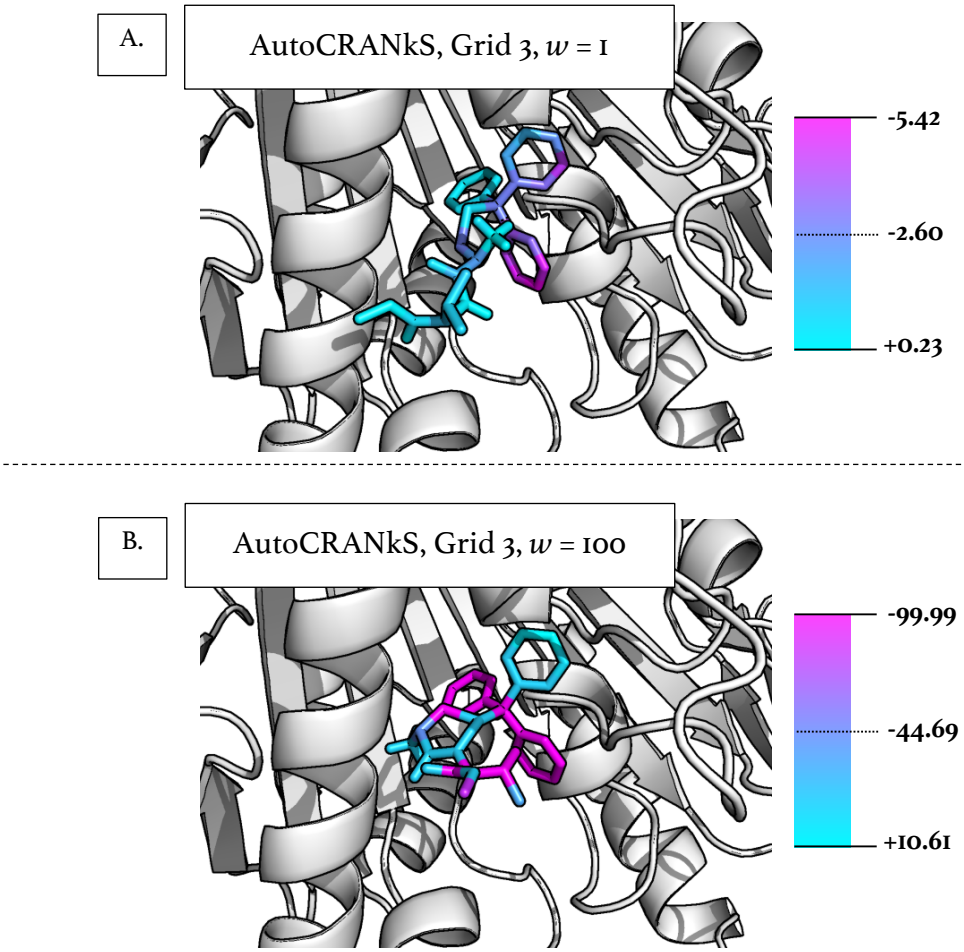



Figure 4.19. Docked poses for active compound 28 from the KIF11 DEKOIS 2.0 dataset. Poses shown are the lowest scored conformer generated using AutoCRANkS with grid 3 and a normalisation constant of 1 (A) and 100 (B). The pose is coloured by the per atom contribution to the score. The 2D structure of the compound is shown in C.

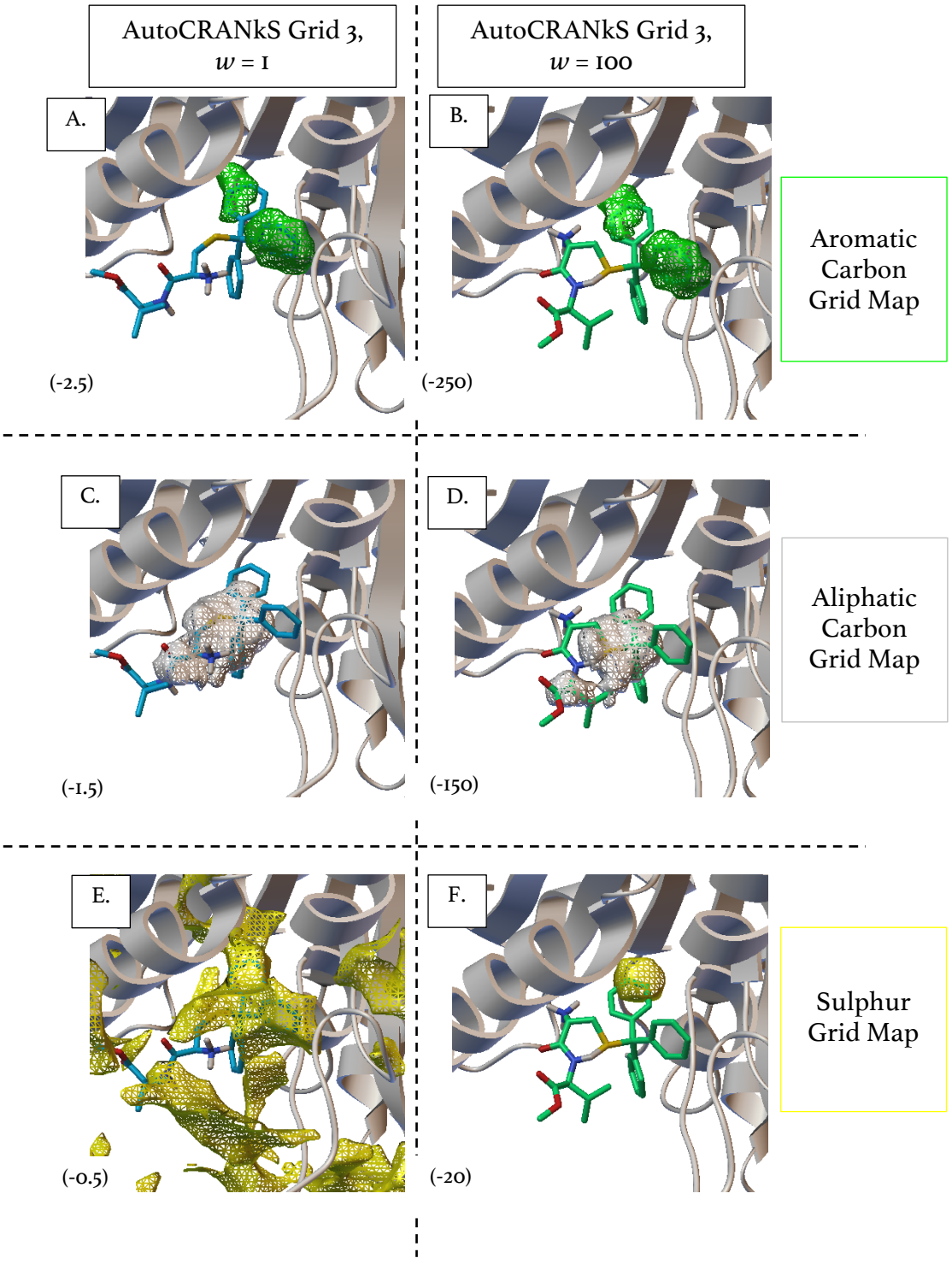



Figure 4.20. Grid maps generated by AutoCRANkS for KIF11 using grid 3 with normalisation constants of 1 and 100. Aromatic carbon grid maps are shown in A ($w$ = 1) and B ($w$ = 100). Aliphatic carbon grid maps are shown in C ($w$ = 1) and D ($w$ = 100). Sulphur grid maps are shown in E ($w$ = 1) and F ($w$ = 100). The isocontour level (in kcal/mol) is shown in brackets. Poses from Figure 4.19 are also shown to allow visualisation of overlap of the pose with the grid maps.

#### *4.3.1.2.4 BCL2*

There is a significant reduction in AUC, $EF_{0.5\%}$ and BEDROC ($\alpha$ = 80.5) for BCL2 when using AutoCRANkS, particularly as the normalisation constant increases above 1 compared to AutoDock. The docking results using the original AutoDock protocol yields an AUC of 0.85 whereas the results when using AutoCRANkS yielded an AUC ranging from 0.42 to 0.43 for a normalisation constant of above 1. AutoDock Vina, GOLD, Glide and AutoDock achieved a high $EF_{0.5\%}$ for this target (between 5 and 20), whereas for AutoCRANkS with a normalisation constant of 1 or above there was zero enrichment.

The distribution of scores using AutoDock and AutoCRANkS using grid 2 are shown in Figure 4.21. For AutoDock and AutoCRANkS with a normalisation constant of 0.01 the distribution is approximately normal. As the normalisation constant increases however the distribution splits into 3 clear peaks. The AUC values are slightly worse than random indicating that this splitting of the distribution does not aid the discrimination between actives and decoys.

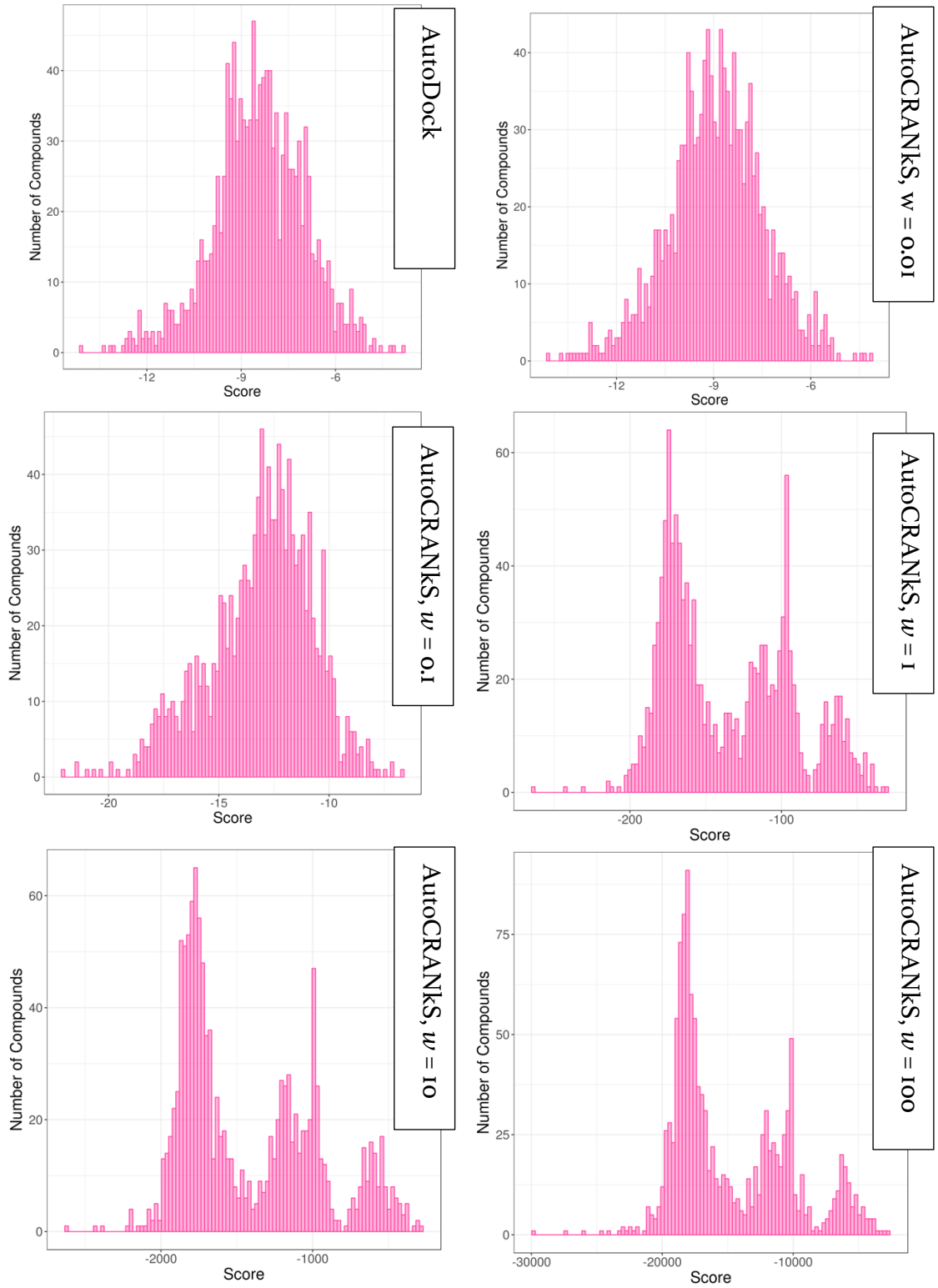



Figure 4.21. Distribution of scores for the actives and decoys from the BCL2 DEKOIS dataset. The compounds were docked by AutoDock and AutoCRANkS using grid 2. As the normalisation constant increases the distribution splits from an approximately normal distribution into 3 clear peaks.

An example of a docked active by AutoDock and AutoCRANkS for normalisation constants 1 is shown in Figure 4.22. This is active compound 26 from the DEKOIS 2.0 dataset for BCL2. For the pose generated by AutoDock the long compound is bent round into a U-shaped pose with carbon atoms and aromatic rings contributing to most of the score (Figure 4.22 A). However, for the pose generated by AutoCRANkS the molecule has extended out (Figure 4.22 B).

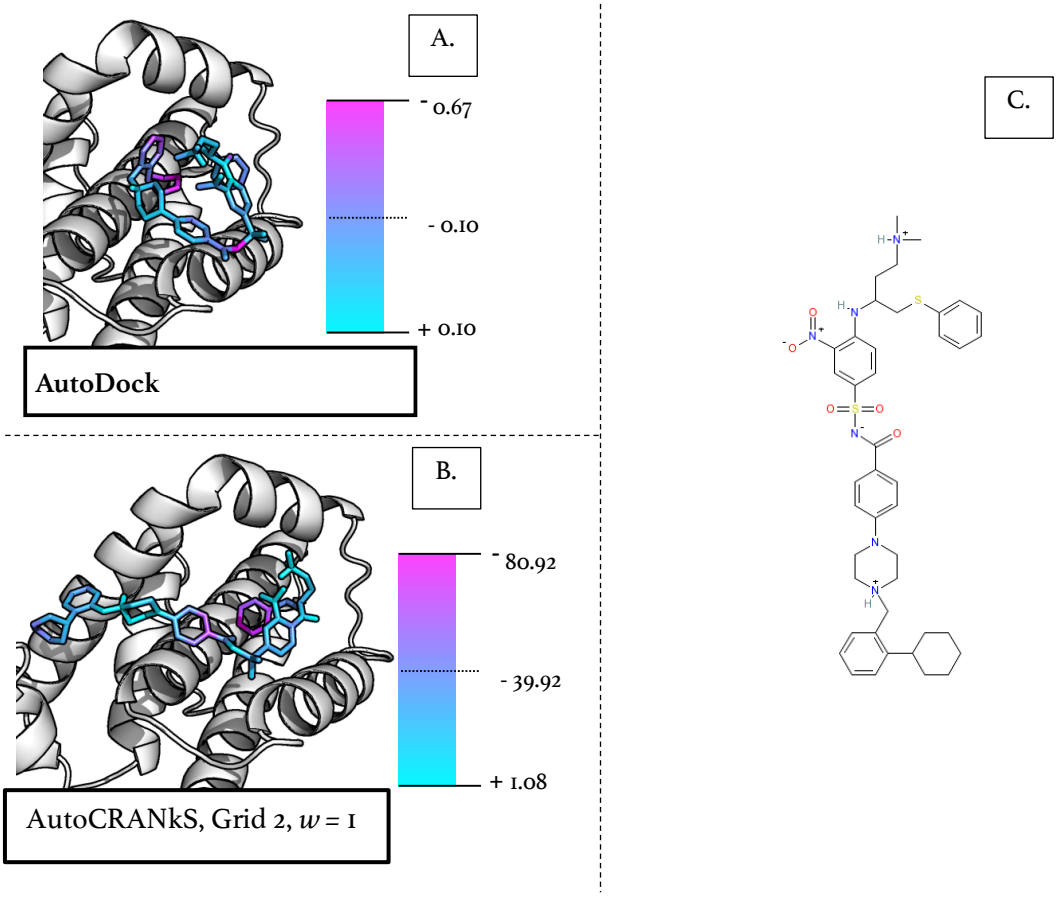



Figure 4.22. Docked poses by AutoDock (A) and AutoCRANkS (B) for active compound 26 from the BCL2 dataset. The poses are coloured by the per atom contribution to the score. The 2D structure of the compound is shown in C. The poses generated are extremely different by each method.

The corresponding grid maps are shown in Figure 4.23. For the AutoDock generated pose the map shows the affinity adopts a ring-link shape around an area of repulsion and the compound fits in around that shape (Figure 4.23 A and C). For a normalisation constant of 1 this shape is lost, and the map is dominated by other pockets of attractive potential introduced by the active grids (Figure 4.23 B and D). The molecule extends out when docked by AutoCRANkS to overlap with these attractive wells.

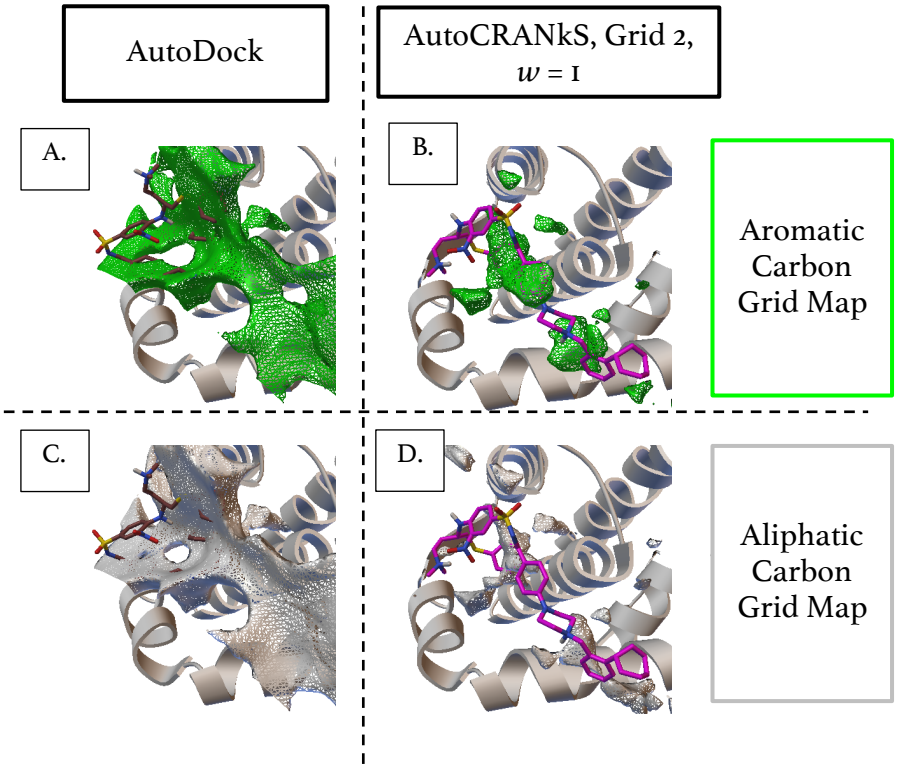



Figure 4.23. Grid maps generated for AutoDock and AutoCRANkS for BCL2. Docked poses of the active compound 26 from the BCL2 DEKOIS 2.0 dataset are also shown. Aromatic carbon grid maps generated by AutoDock (A) and AutoCRANkS (B) are shown. Aliphatic carbon grid maps generated by AutoDock (C) and AutoCRANkS (D) are shown. The attractive potential added by AutoCRANkS causes the pose generated to be long. In comparison the pose generated by AutoDock is bent round to fill the attractive potential around a hole of repulsion present in both A and C.

Figure 4.24 shows the ligands used in the active grid for grid 2 for BCL2. Like the actives and decoys, all the ligands are extremely long, extending out in an orientation similar to the pose generated for the active compound by AutoCRANkS. It is noticeable from Figure 4.24 that one of the alpha helices in the binding site exhibits changes in conformation between the PDB structures and this is labelled. In Chapter 3 I found that the PDB structures that did not all adopt the same conformation achieved poor results for the CRANkS algorithm and it is likely the same behaviour occurs here. The flexible helix forms part of the binding site and this change in position will affect how the ligands overlap, adding noise to the grids and a detrimental effect on performance.

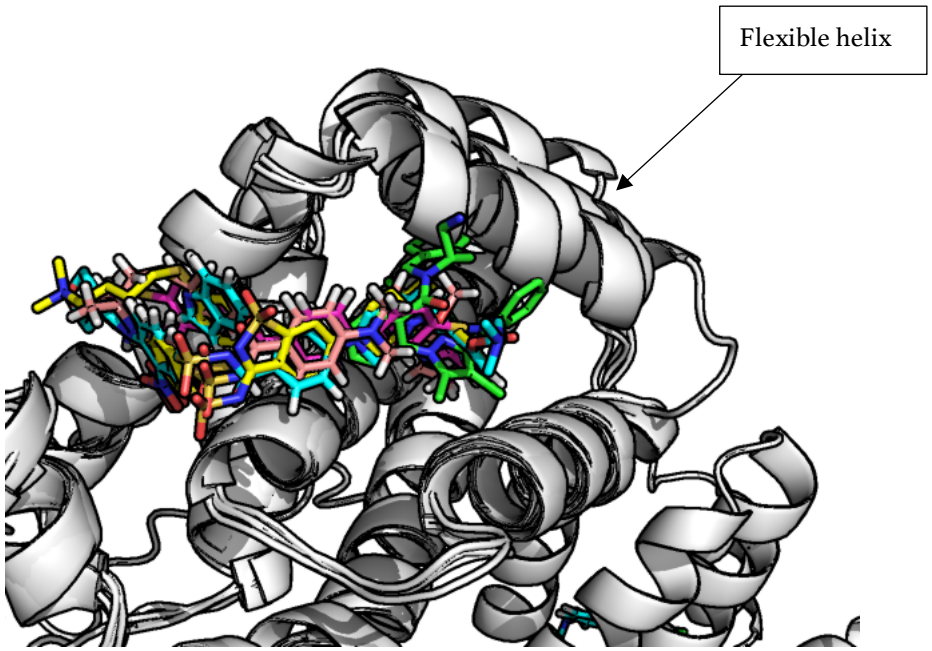



Figure 4.24. Ligands used to construct grid 2 for target BCL2. One helix can be seen to flexible and adopts a different location for different structures – labelled flexible helix.

#### *4.3.1.2.5 Selection of Active Ligands for use in CRANkS Grids*

Targets HSP90A, ADRB2 and DHFR were investigated in Chapter 3 and consequently the same grids used in Chapter 3 were used in this work. These consist of five randomly selected ligands. This should allow comparison between the Chapters to determine whether grids that yielded high performance for the CRANkS algorithm also yield high performance when used in AutoCRANkS. However, few similarities in behaviour were found. For example, for HSP90A when using the CRANkS algorithm, a maximum AUC of 0.94 was achieved when using grid 1 and the lowest AUC of 0.67 was achieved when using grid 3. In comparison the highest AUCs achieved by AutoCRANkS were found using grid 4 (an AUC of 0.71 using a normalisation constant of 1).

The remaining targets used ligands selected for grids by 2D similarity. For target KIF11 all compounds were found to form a single cluster when clustered by 2D similarity and were randomly selected. There is little difference between the best and worst results for this target. For target BCL2 grids 1 to 3 are formed of compounds that formed a single cluster whereas grids 4 and 5 consist of randomly selected ligands. However, there is little difference in performance between the grids despite the difference in molecular similarity. Finally, for target AKT1, grid 1 used ligands from a single cluster and the remaining grids used ligands from separate clusters. However once again there is little difference in performance. For example, for a normalisation constant of 1 the AUC calculated when using grid 1 was 0.62 and when using grid 2 was 0.63.

Consequently, further work could be done to determine the optimal method of selecting ligands for use in the grids. Metrics involving the shape of ligands, overlap between ligands or even the number of heteroatoms could be used to determine relationships between the choice of ligands and the performance of AutoCRANkS.

#### *4.3.1.2.6 AutoCRANkS Summary*

To summarise, I have performed docking of DEKOIS 2.0 sets of actives and decoys for six different targets using AutoDock and AutoCRANkS. The docking performed by AutoCRANkS used five different grids for five different normalisation constants. The performance of the docking in discriminating between actives and decoys was measured using AUC, $EF_{0.5\%}$ and BEDROC ($\alpha$ = 80.5).

For four out of the six targets (HSP90A, ADRB2, DHFR and KIF11) there was a significant increase in AUC when using AutoCRANkS compared to AutoDock, in particular for a normalisation constant of 1. For three of these (ADRB2, DHFR and KIF11) there was also a significant increase in enrichment measured by $EF_{0.5\%}$ and BEDROC ($\alpha$ = 80.5). For these targets when using a normalisation constant of 1 AutoCRANkS was calculated to have a higher $EF_{0.5\%}$ than all docking tools compared to in this study (AutoDock, AutoDock Vina, Glide and GOLD). For HSP90A there was no enrichment for actives when using AutoCRANkS as measured by $EF_{0.5\%}$, however only one docking tool achieved a non-zero enrichment for this target. This also indicates that the improvement in AUC is likely to be due to the down-ranking of decoys, rather than up-ranking of active compounds.

For target AKT1 there is a slight decrease in AUC on using AutoCRANkS, but for the other two metrics there is no difference as AutoDock also achieved zero enrichment. For target BCL2, there is a large decrease in all metrics on using AutoCRANkS compared to AutoDock and the other three docking tools, in particular for high normalisation constants. This is thought to be due to motion of one of a helix in the binding site, as multiple protein conformations were found to have a detrimental effect on the performance of CRANkS in Chapter 3.

Overall these results are promising with AutoCRANkS achieving better discrimination between actives and decoys than AutoDock for four out of six targets and higher enrichment than four widely used docking tools for half the targets. Of the remaining two targets only one target performed significantly worse than other docking tools and AutoDock.

### *4.3.2 Interaction-Filtered AutoCRANkS*

I developed a second method of combining the CRANkS grids with the AutoGrid maps: AutoCRANkS Int. This method only combines the signal from atoms that are calculated to form interactions with the target, with the AutoGrid maps. This filtering could remove noise by only adding atoms calculated to be involved in the binding. To investigate the performance of interaction-filtered AutoCRANkS (AutoCRANkS Int) actives and decoys from the DEKOIS 2.0 set for the same set of six diverse targets were docked. Again, five grids consisting of protein-ligand structures were used for each target. For each of these targets the docking was run

using the five normalisation constants. The results are compared to that of AutoDock and AutoCRANkS using the AUC, $EF_{0.5\%}$ (0,5%) and BEDROC ($\alpha$ = 80.5).

Figure 4.25 shows the AUC calculated for the ranking of actives and decoys for the DEKOIS targets using AutoCRANkS Int. For targets that performed well using AutoCRANkS (HSP90A, ADRB2, DHFR, KIF11) a reduction in AUC can be observed when AutoCRANkS Int is used, shown by the arrows and data bars in Figure 4.25. The data bars show that the difference is small. However, for these targets the AUC remains higher than for AutoDock (for example for DHFR the AUC calculated for AutoDock was 0.47 whereas the AUC calculated for AutoCRANkS Int on average over the grids with a normalisation constant of 1 was 0.73) .

For BCL2 there is still no improvement in AUC between AutoCRANkS Int and AutoDock. The filtering of the maps causes the active grids to have less of an impact on the docking and consequently there is a less negative impact on the AUC. By removing atoms, the contribution is reduced and consequently the effect on performance is reduced. For example, on average for a normalisation constant of 1 the AUC calculated from results generated by AutoCRANkS Int was 0.83 compared to 0.42 calculated for AutoCRANkS and 0.85 for AutoDock. For AKT1 however, there is now an improvement in AUC compared to AutoDock, especially when the best results are considered. For example, when using grid 4 and a normalisation constant of 100 an AUC of 0.78 is calculated for AutoCRANkS Int, compared to 0.62 for AutoCRANkS and 0.67 for AutoDock.  This is interesting as AutoCRANkS achieved lower AUCs than AutoDock for this target indicating that the interaction filtering may be filtering out noise for this target. This is discussed in the following pages.

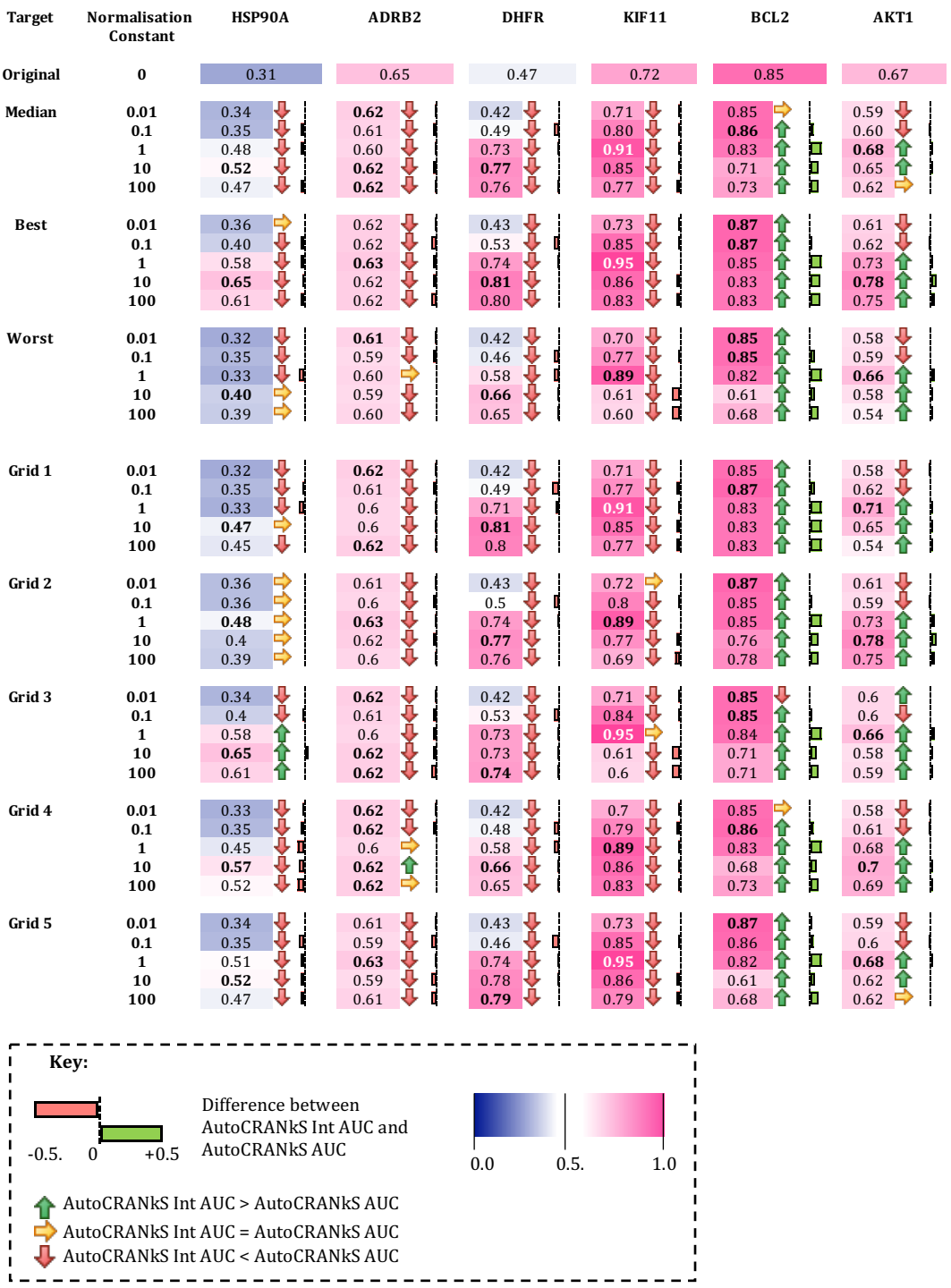



Figure 4.25. AUC for docking of the DEKOIS dataset using AutoCRANkS Int. The median, the highest AUC and the lowest AUC across the five grids for each normalisation constant are shown. Arrows and bars also indicate the difference between the AUC calculated for AutoCRANkS Int and AutoCRANkS. For HSP90A, DHFR and KIF11 for a normalisation constant of 1 and above there is an increase in AUC compared to AutoDock on average but a reduced AUC compared to AutoCRANkS. For AKT1 and ADRB2 there is no significant different between the results for AutoCRANkS Int and AutoDock. For BCL2 there is a decrease in AUC on increasing normalisation constant which is less than for AutoCRANkS.

The $EF_{0.5\%}$ for each target is shown in Figure 4.26 for docking by AutoCRANkS Int. For HSP90A there is little difference between AutoCRANkS Int and AutoCRANkS with on average zero enrichment for all normalisation constants. For the three targets that AutoCRANkS outperformed all other docking tools (DHFR, ADRB2 and KIF11) in terms of enrichment, AutoCRANkS Int achieves similar $EF_{0.5\%}$ values on average to AutoCRANkS. However, AutoCRANkS Int achieves lower enrichment than AutoCRANkS on average at higher normalisation constants for these targets. For example, for KIF11 the enrichment is zero on average for normalisation constants of 10 and 100 when using AutoCRANkS Int but the enrichment achieved for these normalisation constants when using AutoCRANkS were 26 and 5 respectively.

For AKT1 the improved performance is not seen in enrichment at this level. The median across the grids for the enrichment achieved when using AutoCRANkS Int is still zero. This is the same as achieved by AutoCRANkS and AutoDock. However, for BCL2 there is increased enrichment when using AutoCRANkS Int compared to AutoCRANkS, shown by the green arrows in Figure 4.26. For example, the enrichment on average when using a normalisation constant of 1 and AutoCRANkS Int is 10 compared to a median enrichment of zero for a normalisation constant of 1 when using AutoCRANkS. However, for no normalisation constants is the median $EF_{0.5\%}$ across the grids higher than for AutoDock for this target.

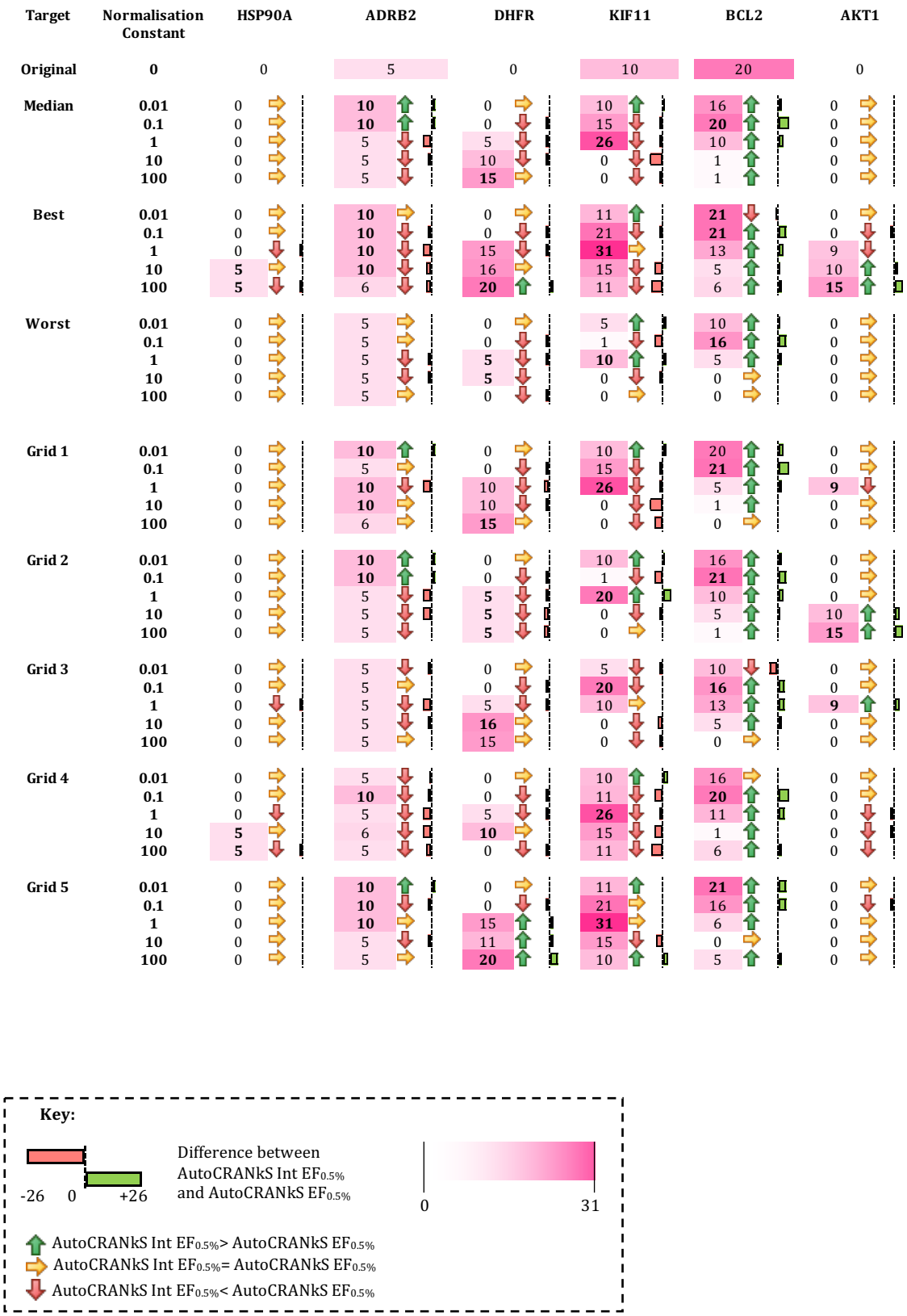



Figure 4.26. $EF_{0.5\%}$ for docked DEKOIS datasets using AutoCRANkS Int. The colour of each cell indicates the enrichment. Arrows and bars indicate the difference between the enrichment calculated for AutoCRANkS Int and the enrichment calculated for AutoCRANkS. Only KIF11 and DHFR achieve higher enrichment than AutoDock for a normalisation constant of 1.

Figure 4.27 shows the calculated BEDROC (α = 80.5) for dockings using AutoCRANkS Int. For ADRB2, DHFR and KIF11 the BEDROC (α = 80.5) is higher for the average across the five grids when using AutoCRANkS Int than when using AutoDock. For example, for a normalisation constant of 1 the median of the BEDROC (α = 80.5) across the grids when using AutoCRANkS Int for ADRB2 is 0.13, for DHFR is 0.19 and for KIF11 is 0.67, compared to 0.14, 0.02 and 0.20 achieved by AutoDock for each target respectively. However, the values are slightly lower than for AutoCRANkS both on average across the grids and for the best results. For example, using a normalisation constant of 1 the median BEDROC (α = 80.5) using AutoCRANkS is 0.40 for ADRB2, 0.27 for DHFR and 0.79 for KIF11, so the difference between AutoCRANkS and AutoCRANkS Int for each target is 0.27, 0.08 and 0.15 respectively.

The BEDROC values for HSP90A are very low for both AutoCRANkS Int and AutoCRANkS, with zero BEDROC achieved in nearly all cases. For AKT1 there is a small improvement on average for AutoCRANkS Int compared to AutoCRANkS. However, there is a notable improvement for one grid (grid 2) when using AutoCRANkS Int and the best BEDROC (α = 80.5) values for normalisation constants of 10 and 100 are significantly higher. For BCL2, there is a big improvement in the BEDROC (α = 80.5) compared to AutoDock and AutoCRANkS in particular for a normalisation constant of 1. Even for high normalisation constants there are high BEDROC (α = 80.5) values when using AutoCRANkS Int for this target, which was not achieved when using AutoCRANkS.

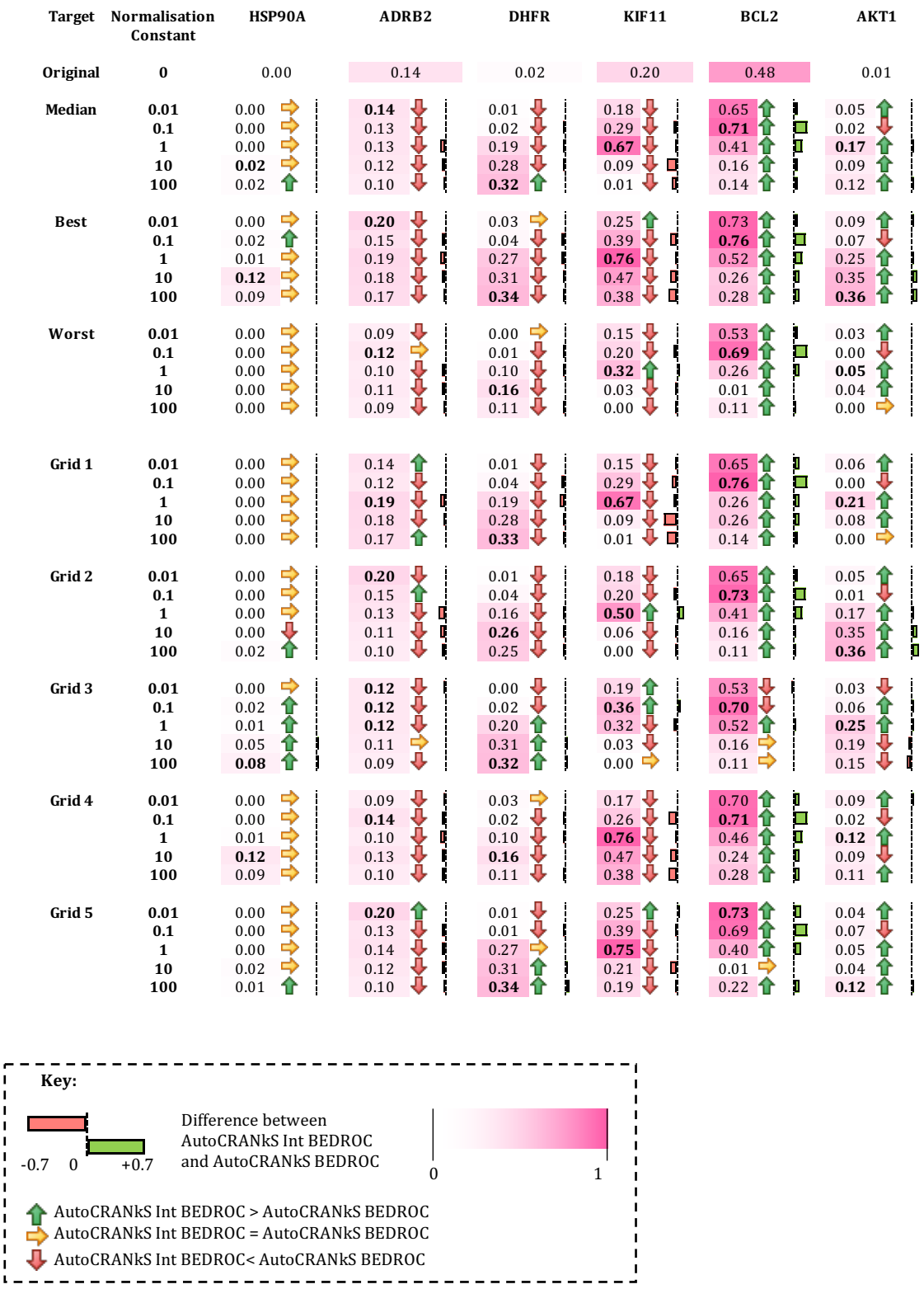



Figure 4.27. BEDROC (α = 80.5) for docked results for DEKOIS datasets using AutoCRANkS Int. For DHFR, KIF11 And AKT1 there is a significant increase over AutoDock using a normalisation constant of 1. Arrows and data bars indicate the difference between the BEDROC calculated for AutoCRANkS Int and the BEDROC calculated for AutoCRANkS.

#### *4.3.2.1 DHFR*

Similar to results when using AutoCRANkS, AutoCRANkS Int showed a large improvement over using AutoDock in terms of higher AUCs, $EF_{0.5\%}$ and BEDROC ($\alpha$ = 80.5) for this target (see Figures 4.25 to 4.27). In particular results for AutoCRANkS Int are best for high normalisation constants of 10 and 100 (on average AUC of 0.77 and 0.76 respectively across the five grids), and this is mirrored in the performance of AutoCRANkS (an average AUC of 0.78 and 0.79 respectively).

Figure 4.28 shows the distribution of the scores for the docking of actives and decoys for the DHFR dataset using AutoCRANkS Int for grid 1 for three normalisation constants alongside the distributions for AutoCRANkS for the same normalisation constants. As the normalisation constant increases to 1 and above the distribution splits into two peaks with a smaller peak at high scores. This is the same behaviour exhibited by AutoCRANkS and the interaction filtering has had no large impact on the distribution.

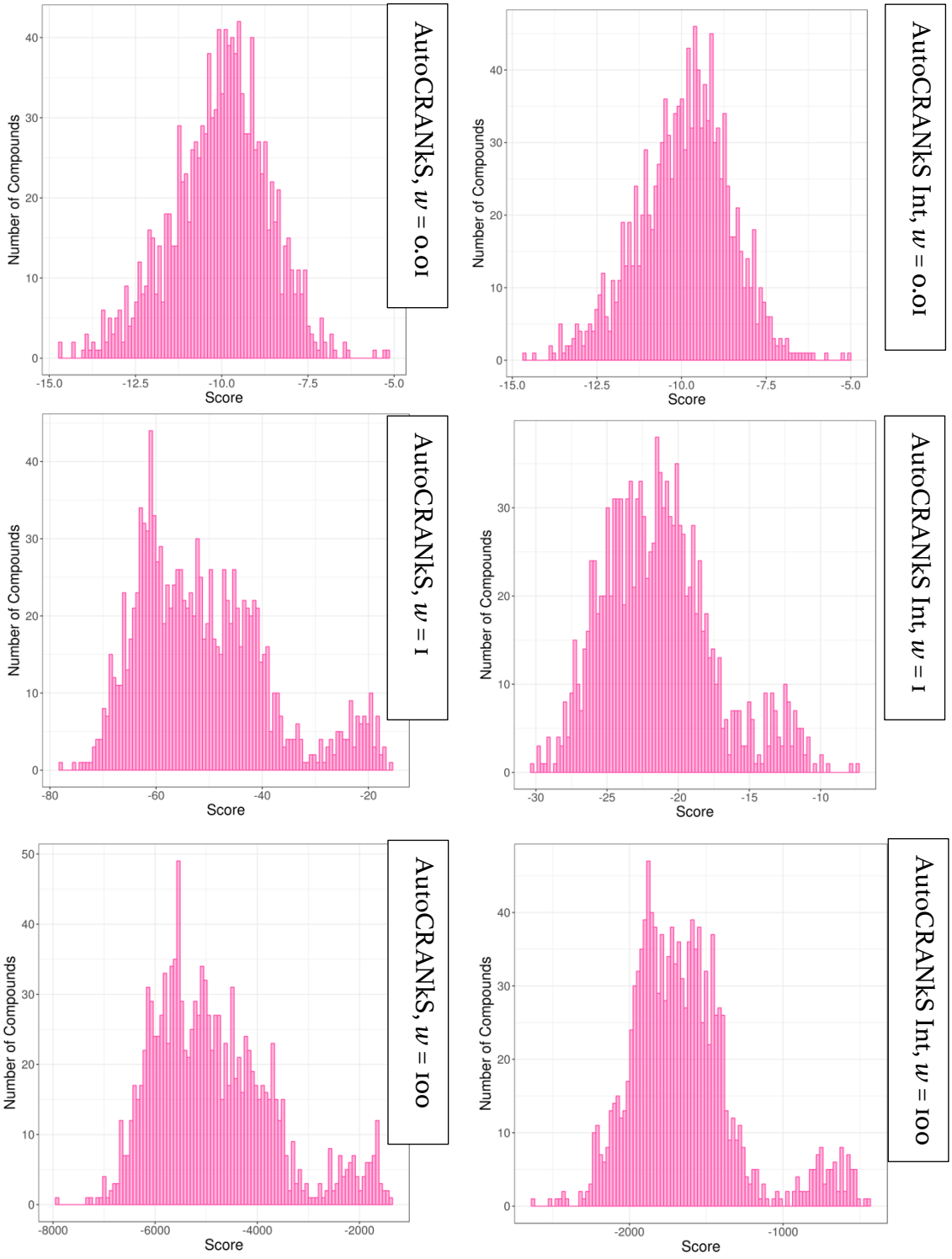



Figure 4.28, Distribution of scores for the DHFR dataset docked by AutoCRANkS and AutoCRANkS Int using grid 1. There is no clear difference between these distributions.

Examples of a docked pose generated by AutoDock Int is shown in Figure 4.29 alongside a pose generated by AutoCRANkS. This is for active compound 9 from the DEKOIS 2.0 DHFR dataset and the 2D structure is shown in Figure 4.29 C. For this active compound using grid 1 with a normalisation constant of 1 AutoCRANkS Int ranked the compound 10$^{th}$ and AutoCRANkS ranked the compound 5$^{th}$. The poses are being scored similarly with respect to the rest of the set.

The poses generated using grid 1 are similar between AutoCRANkS and AutoCRANkS Int (Figure 4.29 A and B). The main different is a change in the conformation of the diol tail. However, the scoring pattern is different, with the aromatic rings making less of a significant contribution to the score for AutoCRANkS Int (Figure 4.29 B). In this case the main contribution is from an oxygen atom of the diol tail and from a carbon atom on a branch below the fused aromatic rings (Figure 4.29 B).

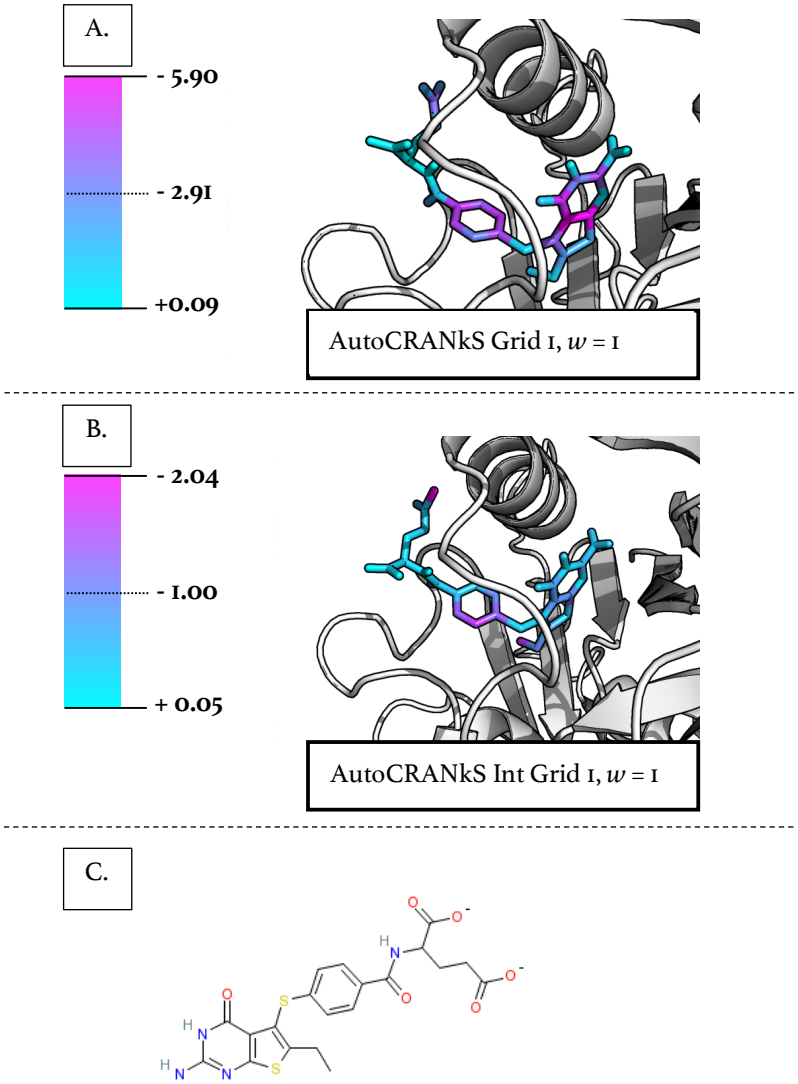



Figure 4.29. Docked pose for active compound 9 from the DHFR dataset generated by AutoCRANkS (A) and AutoCRANkS Int (B) using grid 1 and a normalisation constant of 1. The 2D structure of the compound is shown in C. The poses are coloured by the per atom contribution to the score. Although the poses are similar the scoring pattern is different with the AutoCRANkS scoring dominated by aromatic atoms. The AutoCRANkS Int scoring is dominated by oxygen atoms.

The corresponding grid maps are shown in Figure 4.30. A similar pattern for the aromatic map is observed between AutoCRANkS (Figure 4.30 A) and AutoCRANkS Int (Figure 4.30 B). However, a higher isocontour level is needed for the AutoCRANkS Int map to achieve this pattern. This explains why the score of the aromatic rings is less dominant when scored by AutoCRANkS Int. The aliphatic carbon map shows a difference between the AutoCRANkS (Figure 4.30 C) and AutoCRANkS Int (Figure 4.30 D) maps. The well of attractive potential of the AutoCRANkS map that keeps the diol branch in place (Figure 4.30 C) was from an atom that was not calculated to form any interactions, as this attractive well is not present in the interaction filtered map (Figure 4.30 D). The oxygen acceptor map shows that for both methods the oxygen atoms of the diol group occupy the same area of attractive potential (Figure 4.30 E and F). However, as the AutoCRANkS Int scoring is not so dominated by the aromatic map, the oxygen map makes a larger contribution to the score.

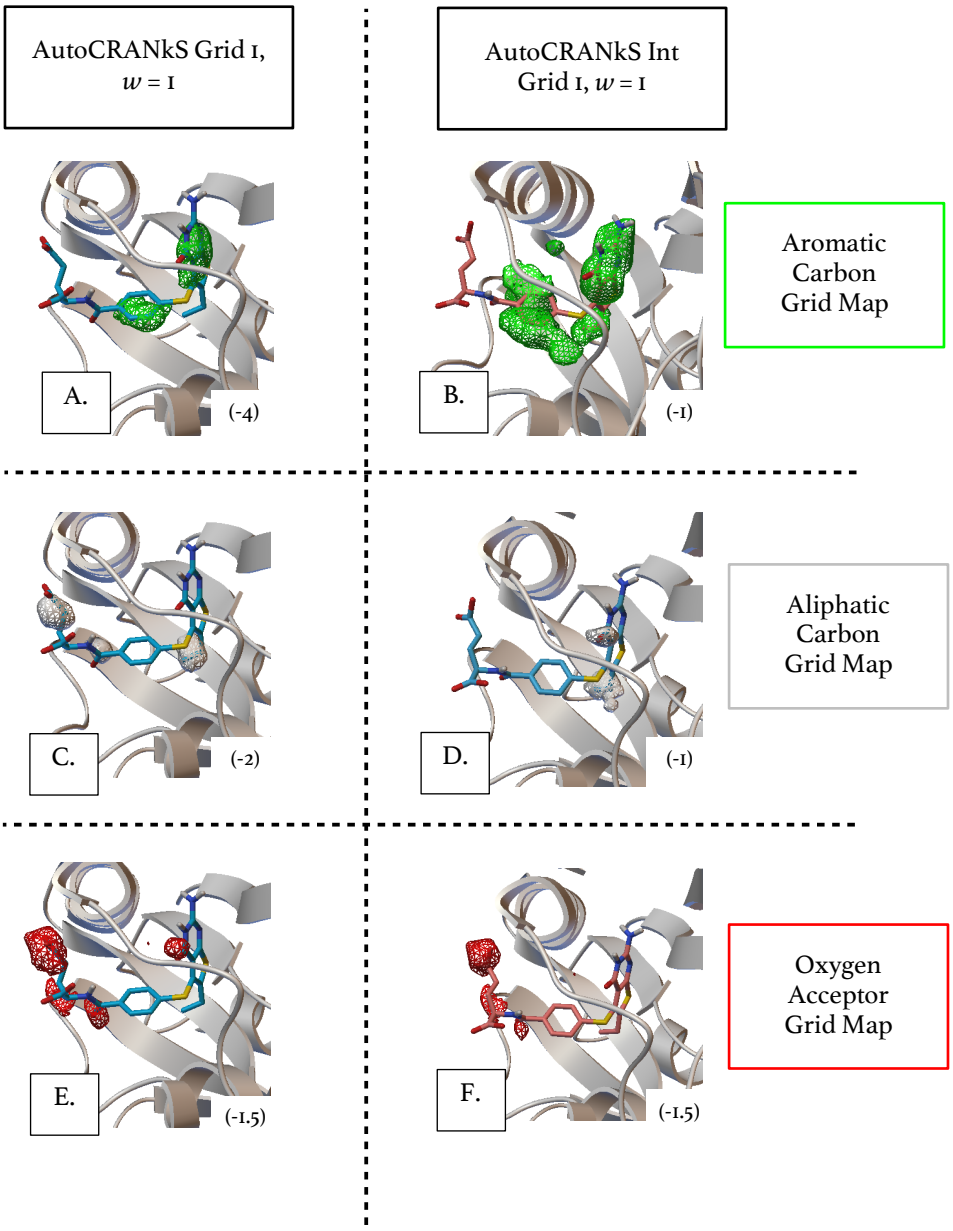



Figure 4.30. Docked poses for active compound 9 from the DEKOIS 2.0 DHFR dataset generated by AutoCRANkS and AutoCRANkS Int using grid 1 and normalisation constant 1. Aromatic carbon grid maps are shown generated by AutoCRANkS (A) and AutoCRANkS Int (B). Aliphatic carbon grid maps are shown generated by AutoCRANkS (C) and AutoCRANkS Int (D). Oxygen acceptor grid maps are shown generated by AutoCRANkS (E) and AutoCRANkS Int (F). The isocontour level in kcal/mol is shown in brackets for each grid. Higher isocontour levels are required to see attractive wells in the aromatic grid map for AutoCRANkS Int (B) than AutoCRANkS (A) and so there is less contribution from the aromatic map to the overall score.

Table 4.5 shows the number of atoms used in the active grid by AutoCRANkS and AutoCRANkS Int for grid 1. A reduction in the number of aromatic carbon and aliphatic carbon atoms is clear. This indicates that heteroatom grids are more likely to contribute to the score for the AutoCRANkS Int method, as determined by inspection of the grid maps in Figure 4.30.

| AutoDock Atom Type | AutoCRANkS | AutoCRANkS Int |
|---|---|---|
| Aromatic Carbon | 88 | 20 |
| Aliphatic Carbon | 31 | 6 |
| Nitrogen Acceptor | 18 | 8 |
| Oxygen Acceptor | 14 | 8 |

Table 4.5. Number of atoms in the active grid for grid 1 for target DHFR that will be added to the AutoGrid maps.

Another example of docked poses generated by AutoCRANkS and AutoCRANkS Int for an active compound is shown in Figure 4.31. This active compound is active compound 2 from the DEKOIS 2.0 dataset for DHFR and the 2D structure is shown in Figure 4.31 C. It was ranked 21$^{st}$ by AutoCRANkS and 246$^{th}$ by AutoCRANkS Int. In this case the poses generated are dissimilar (Figure 4.31 A and B). The scoring pattern indicates that the aromatic potential is dominating both scores as the poses show the lowest scores over the aromatic rings in both cases.

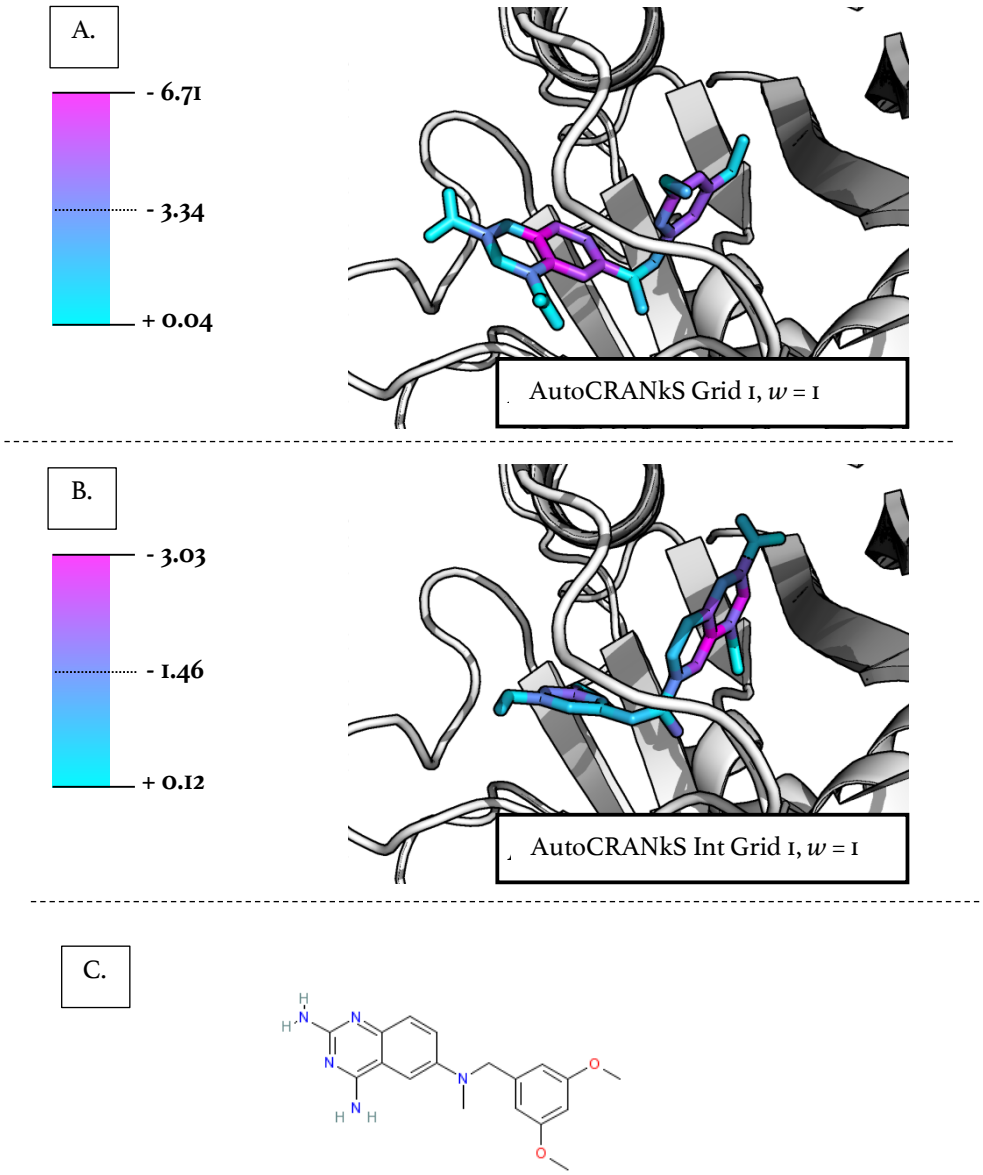



Figure 4.31. Docked poses for active compound 2 from the DEKOIS 2.0 DHFR dataset using AutoCRANkS (A) and AutoCRANkS Int (B) using grid 1 and a normalisation constant of 1. The pose is coloured by the per atom contribution to the score in each case. The 2D structure of the compound is shown in C. The poses generated are different by each method however both scores are dominated by aromatic carbon contributions.

The corresponding grid maps are shown in Figure 4.32. For the AutoCRANkS grid map there are two clear wells of attractive potential in the aromatic carbon grid that have been introduced by the active grids (Figure 4.32 A). Aromatic rings occupy both areas leading to low scores for both groups. In contrast for the AutoCRANkS Int grid map a higher isocontour level is required for both aromatic groups to be covered (Figure 4.32 B). Even using the higher isocontour level the benzyl ring is only partially covered, and this can be seen in Figure 4.31 B as only two of atoms of this aromatic ring are low scoring. Some atoms from the active grid that can be seen in the AutoCRANkS aromatic map clearly do not form interactions and have been eliminated from the AutoCRANkS Int map (Figure 4.32 B). The conserved aromatic potential on the right of the figure also looks reduced for the AutoCRANkS Int grid map indicating there are a reduction of atoms contributing to the active grid in this area, due to a lack of interactions (Figure 4.32 B).

The aliphatic carbon grid map also shows a reduction in attractive wells of potential caused by active grid atoms (Figure 4.32 D). For the AutoCRANkS Int map a high isocontour level is needed to show the overlap between the methyl of the tertiary amine with attractive potential, indicating this potential is not from an active grid (Figure 4.32 D). The aliphatic grid map for AutoCRANkS shows a different pattern (Figure 4.32 C). The majority of the active grid signal present in the AutoCRANkS aliphatic carbon grid map is eliminated in the AutoCRANkS Int map due to a lack of interactions. The methyl group of the tertiary amine is again held in a well of attractive potential in the AutoCRANkS map (Figure 4.32 C).

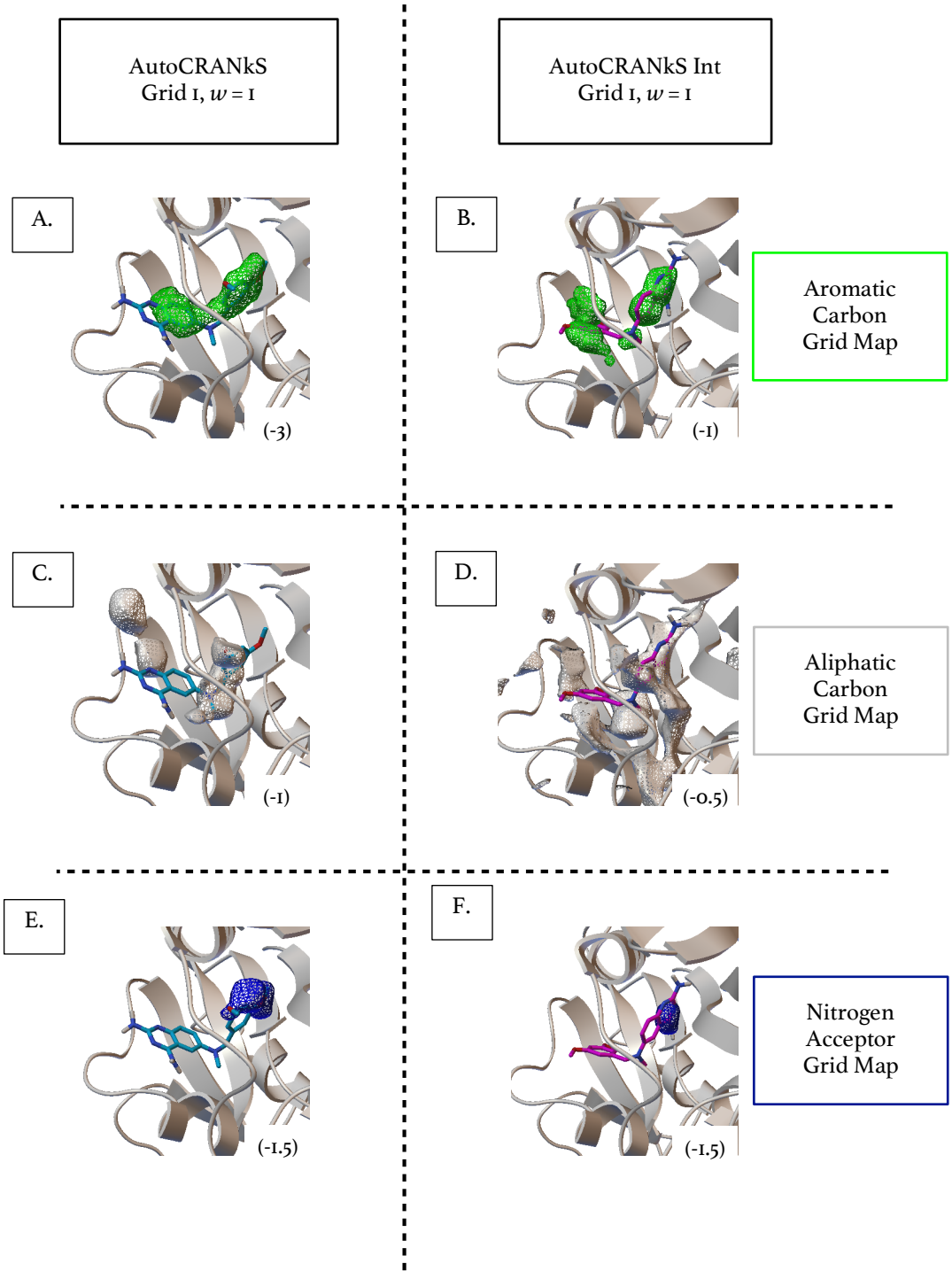



Figure 4.32. Docked poses for active compound 2 from the DHFR DEKOIS 2.0 dataset using AutoCRANkS and AutoCRANkS Int with grid 1 and a normalisation constant of 1. Aromatic carbon grid maps are shown generated by AutoCRANkS (A) and AutoCRANkS Int (B). Aliphatic carbon grid maps are shown generated by AutoCRANkS (C) and AutoCRANkS Int (D). Nitrogen acceptor grid maps are shown generated by AutoCRANkS (E) and AutoCRANkS Int (F). The pose is coloured by the per atom contribution to the score in each case. The isocontour level used for each map in kcal/mol is shown in brackets.

The nitrogen acceptor grid maps show a similar pattern for both methods (Figure 4.32 E and F), although the reduced shape and span of the AutoCRANkS Int attractive potential indicates a reduction in the nitrogen atoms contributing to this area. The AutoCRANkS pose shows no overlap of a nitrogen with this area (Figure 4.32 E). However, the AutoCRANkS Int pose does overlap a nitrogen with this attractive potential (Figure 4.32 F), and this atom is shown to contribute significantly to the score in Figure 4.31 B.

Overall for this active compound the large attractive aromatic areas present in the AutoCRANkS map allow the active compound to be determined to be active by AutoCRANkS. For AutoCRANkS Int, despite overlap with the nitrogen map, the lack of these areas due to the removal of aromatic atoms from the active grids, causes a more positive score. The active compound is therefore ranked poorly.

### *4.3.2.2 AKT1*

The docking of the AKT1 dataset by AutoCRANkS showed a small reduction in AUC, $EF_{0.5\%}$ and BEDROC compared to AutoDock. However, when the docking is performed by AutoCRANkS Int there is a small improvement compared to AutoDock, with increases on average for the AUC, $EF_{0.5\%}$ and BEDROC ($\alpha$ = 80.5). In particular using grid 2 with AutoCRANkS Int showed high performance with an AUC of 0.78 for a normalisation constant of 10 (compared to an AUC of 0.67 using AutoDock and an AUC of 0.60 with the same normalisation constant and grid using AutoCRANkS), an $EF_{0.5\%}$ of 10 (compared to 0 using AutoDock and AutoCRANkS)

and a BEDROC (α = 80.5) of 0.35 (compared to 0.01 using AutoDock and 0.07 using AutoCRANkS).

Poses generated by AutoCRANkS Int and AutoCRANkS for active compound 4 from the DEKOIS 2.0 dataset for AKT1 using grid 2 with a normalisation constant of 10 are shown in Figure 4.33. The poses are similar but the aliphatic linker in the centre of the molecule arcs in opposite directions. The AutoCRANkS pose shows the biggest contribution to the score is by atoms in the bottom fused aromatic rings (Figure 4.33 A). The AutoCRANkS Int pose contrastingly shows atoms in both aromatic centres contributing strongly to the score (Figure 4.33 B).

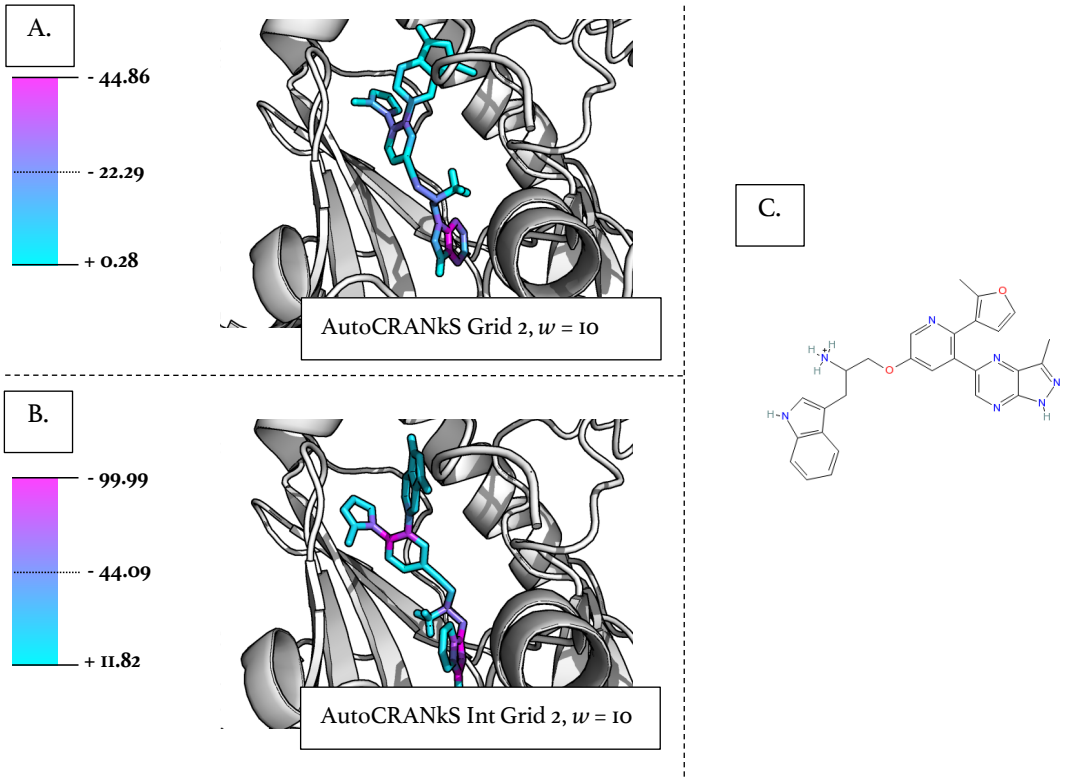



Figure 4.33. Docked poses generated for active compound 4 from the AKT1 DEKOIS 2.0 dataset by AutoCRANkS (A) and AutoCRANkS Int (B) using grid 2 and a normalisation constant of 10. The poses are coloured by each atom's contribution to the score. The 2D structure of the compound is shown in C. For AutoCRANkS the score is dominated by two atoms in the fused aromatic rings at the bottom of A whereas for the pose generated by AutoCRANkS Int both aromatic centres contribute significantly to the score (B)

The corresponding grid maps are shown in Figure 4.34. For this compound only the aromatic and carbon grid maps were found contribute to the score. There is overlap with aromatic potential for both poses (Figure 4.34 A and B). However, the orientation of the pose generated by AutoCRANkS Int better overlaps with the top aromatic area of attractive potential (Figure 4.34 B). If the aliphatic carbon map is considered it can be seen that the aliphatic linker is held in place by a single area of attractive potential in the case of AutoCRANkS Int (Figure 4.34 D). In contrast the AutoCRANkS map has a much broader area of potential in the region allowing the linker to adopt several orientations without holding the molecule in place for maximum overlap with the aromatic potential (Figure 4.34 C).

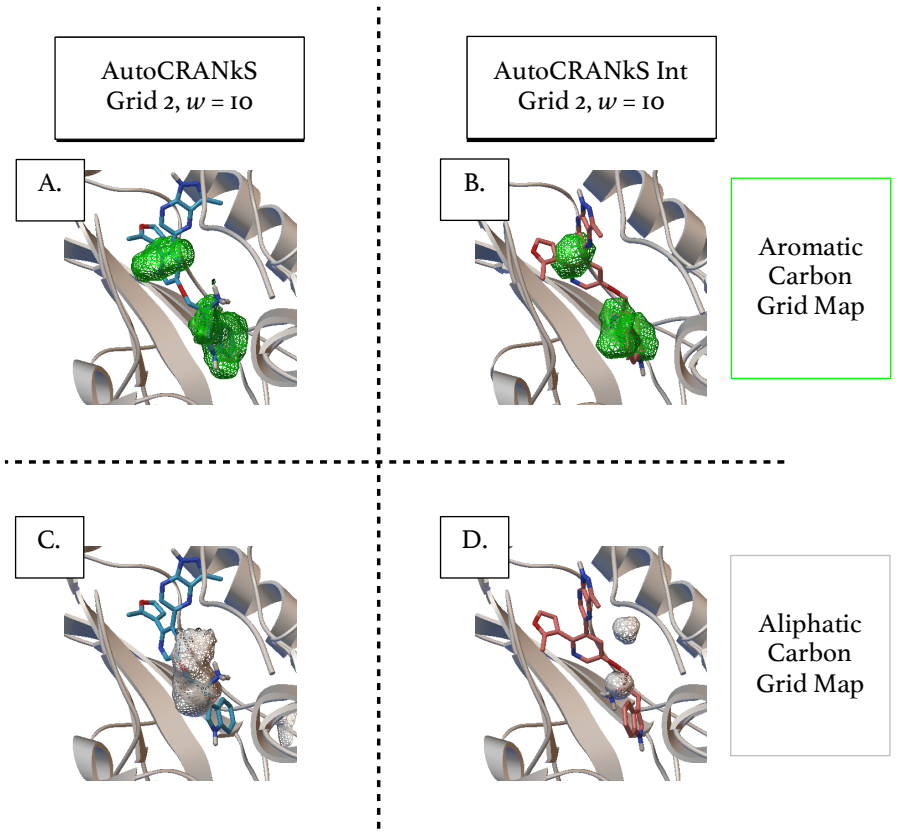



Figure 4.34. Docked poses generated by AutoCRANkS and AutoCRANkS Int for active compound 4 from the AKT1 DEKOIS 2.0 dataset are shown. The aromatic carbon grid maps are shown generated by AutoCRANkS (A) and AutoCRANkS Int (B). The aliphatic carbon grid maps are shown generated by AutoCRANkS (C) and AutoCRANkS Int (D). The large well of attractive potential in aliphatic carbon grid map from AutoCRANkS (A) allows the orientation of the pose to be flexible and not match with both attractive wells in the aromatic grid map (C).

This could indicate that by reducing the areas of attractive potential using the interaction filtering there is a more stringent criteria for molecules to be able to overlap with multiple wells of attractive potential. This could allow better discrimination between actives and decoys. For this target the filtering reduces the noise and means that only areas that are important for interactions and binding are included in the active grids.

### *4.3.2.3 AutoCRANkS Int Summary*

To summarise, the performance of AutoCRANkS Int at discriminating between actives and decoys was generally found to be reduced compared to AutoCRANkS. For targets for which AutoCRANkS performed well (DHFR, KIF11, ADRB2 and HSP90A) AutoCRANkS Int also performed well but achieved lower AUC, $EF_{0.5\%}$ and BEDROC ($\alpha$ = 80.5) values on average. However, these values were still higher than for AutoDock. This was investigated for target DHFR, and the reduction in performance can be attributed to too many atoms being removed from the active grids and consequently a reduction in valuable signal.

For target BCL2, for which AutoCRANkS performed poorly, AutoCRANkS Int achieves better results. The filtering of active grids means less of the signal that is detrimental to AutoCRANkS is added to the grid. For target AKT1, AutoCRANkS achieved slightly lower AUC values than for AutoDock. However, for AutoCRANkS Int on average the AUC is slightly higher. The same is found for the BEDROC ($\alpha$ = 80.5). On inspection this could be attributed to the significant reduction of the

aromatic and carbon atoms in the active grid imposing a more stringent criteria for molecules to be able to overlap with multiple wells of attractive potential.

### *4.3.3 Scoring of AutoCRANkS Poses by AutoDock*

To investigate the effect of the docked poses generated by AutoCRANkS on the discrimination between actives and decoys, the poses are rescored using the AutoDock scoring function with no inclusion of active grids. This will be referred to as AutoDock | AutoCRANkS. The rescoring is calculated for all 10 poses generated by AutoCRANkS for each compound. The AUC, $EF_{0.5\%}$ and BEDROC ($\alpha$ = 80.5) are then calculated for the dataset for each target using the lowest score calculated for one of the 10 poses for each compound.

The AUC values calculated for the datasets using AutoDock | AutoCRANkS are shown in Figure 4.35. For HSP90A the results are similar to AutoCRANkS when averaged across the grids – for a normalisation constant of 1 the difference in AUC between AutoDock | AutoCRANkS and AutoCRANkS is -0.07. However, the best results using AutoDock | AutoCRANkS are worse than for AutoCRANkS: a reduction of -0.17 for a normalisation constant of 1. The $EF_{0.5\%}$, shown in Figure 4.36, show a similar performance of AutoDock | AutoCRANkS to AutoCRANkS with zero enrichment on average for all normalisation constants. This is mirrored by the BEDROC ($\alpha$ = 80.5) where the results are similar for both AutoDock | AutoCRANkS and AutoCRANkS with an average zero enrichment for all normalisation constants (Figure 4.37).

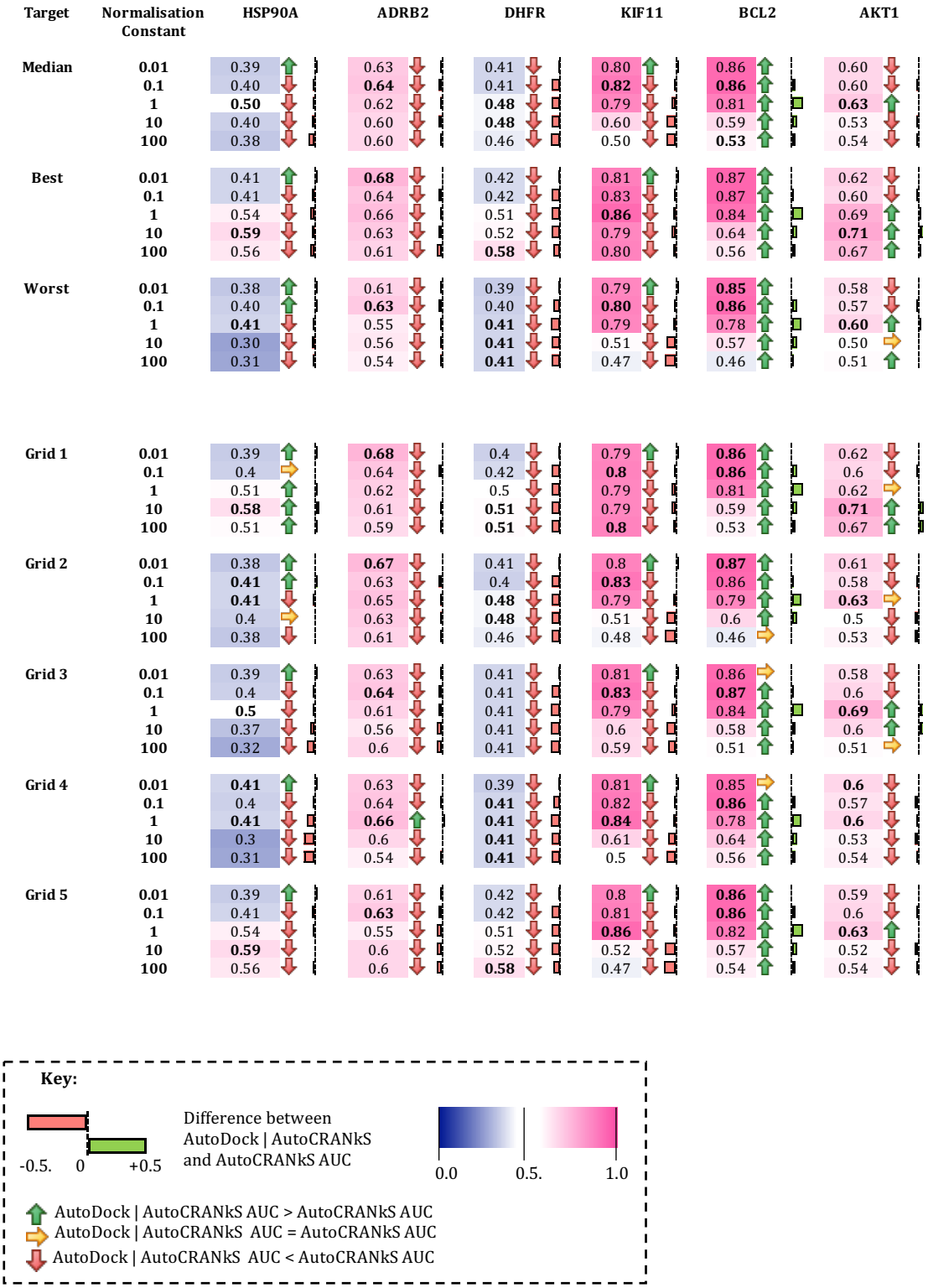



Figure 4.35 AUC values for rescoring of AutoCRANkS docked poses with AutoDock. The median, highest and lowest values across the five grids for each normalisation constant are shown with the colour scaled by the AUC value. The arrows indicate whether the AUC calculated for AutoDock | AutoCRANkS is higher, the same as or lower than for AutoCRANkS. The bars indicate the difference in AUC between AutoDock | AutoCRANkS. For all targets excluding BCL2 the calculated AUCs are worse for AutoDock | AutoCRANkS for normalisation constants > 0.01.

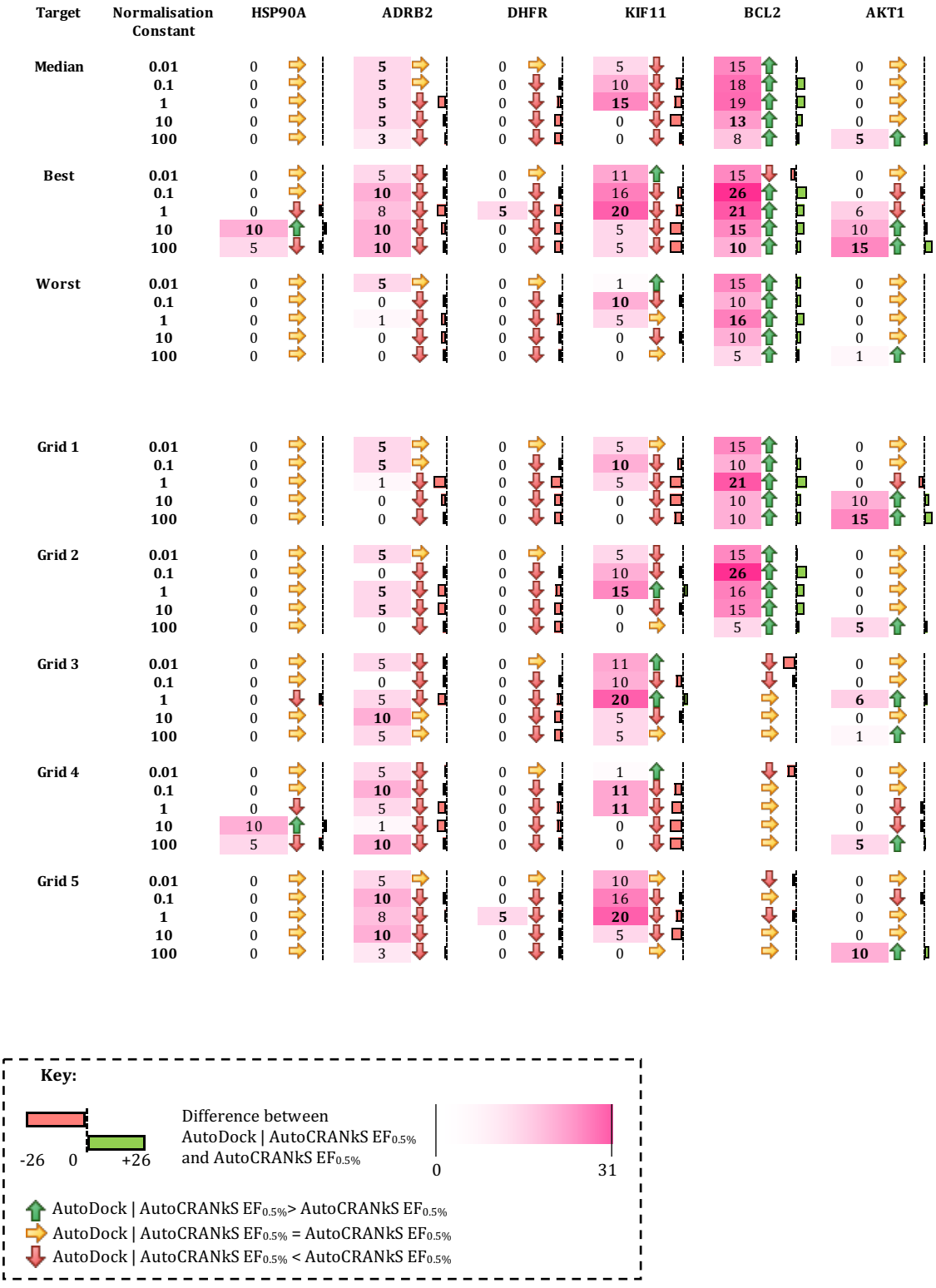



Figure 4.36. $EF_{0.5\%}$ values for rescoring of AutoCRANkS docked poses with AutoDock. The median, highest and lowest values across the five grids for each normalisation constant are shown. The arrows indicate whether the $EF_{0.5\%}$ calculated for AutoDock | AutoCRANkS is higher, the same as or lower than for AutoCRANkS. The bars indicate the difference in $EF_{0.5\%}$ between AutoDock | AutoCRANkS and AutoCRANkS. For all targets excluding BCL2 the calculated $EF_{0.5\%}$ are worse for AutoDock | AutoCRANkS for normalisation constants > 0.01.

| Target | Normalisation Constant | HSP90A | ADRB2 | DHFR | KIF11 | BCL2 | AKT1 |
|---|---|---|---|---|---|---|---|
| Median | 0.01 | 0.00 | 0.13 | 0.01 | 0.24 | 0.49 | 0.03 |
| | 0.1 | 0.00 | **0.14** | 0.01 | 0.31 | 0.52 | 0.02 |
| | 1 | 0.00 | 0.11 | **0.03** | **0.34** | **0.55** | 0.07 |
| | 10 | 0.00 | 0.10 | 0.02 | 0.09 | 0.23 | 0.03 |
| | 100 | 0.00 | 0.11 | 0.01 | 0.02 | 0.11 | **0.11** |
| Best | 0.01 | 0.00 | 0.19 | 0.02 | 0.26 | 0.53 | 0.04 |
| | 0.1 | 0.00 | 0.17 | 0.03 | 0.33 | **0.59** | 0.04 |
| | 1 | 0.00 | 0.15 | **0.06** | **0.44** | 0.58 | 0.19 |
| | 10 | 0.11 | 0.18 | 0.02 | 0.20 | 0.33 | 0.18 |
| | 100 | **0.13** | **0.22** | 0.03 | 0.20 | 0.18 | **0.27** |
| Worst | 0.01 | 0.00 | 0.08 | 0.00 | **0.21** | 0.45 | 0.02 |
| | 0.1 | 0.00 | 0.04 | 0.00 | **0.21** | 0.42 | 0.01 |
| | 1 | 0.00 | 0.08 | 0.00 | 0.17 | **0.37** | 0.04 |
| | 10 | 0.00 | 0.09 | 0.00 | 0.03 | 0.21 | 0.00 |
| | 100 | 0.00 | **0.10** | 0.00 | 0.00 | 0.09 | **0.09** |
| Grid 1 | 0.01 | 0.00 | 0.08 | 0.02 | **0.22** | 0.45 | 0.03 |
| | 0.1 | 0.00 | 0.14 | **0.03** | 0.21 | 0.42 | 0.01 |
| | 1 | 0.00 | **0.15** | 0.04 | 0.17 | **0.55** | 0.07 |
| | 10 | 0.00 | 0.10 | 0.00 | 0.20 | 0.23 | 0.18 |
| | 100 | **0.02** | 0.10 | 0.01 | 0.20 | 0.11 | **0.27** |
| Grid 2 | 0.01 | 0.00 | **0.19** | 0.01 | 0.25 | 0.52 | 0.04 |
| | 0.1 | 0.00 | 0.04 | 0.02 | **0.32** | **0.59** | 0.02 |
| | 1 | 0.00 | 0.09 | **0.03** | 0.31 | 0.58 | 0.07 |
| | 10 | 0.00 | 0.09 | 0.02 | 0.03 | 0.28 | 0.00 |
| | 100 | 0.00 | 0.11 | 0.01 | 0.00 | 0.11 | **0.10** |
| Grid 3 | 0.01 | 0.00 | 0.11 | 0.01 | 0.24 | 0.48 | 0.02 |
| | 0.1 | 0.00 | 0.08 | 0.00 | 0.31 | **0.52** | 0.03 |
| | 1 | 0.00 | 0.08 | 0.00 | **0.42** | 0.41 | **0.19** |
| | 10 | 0.00 | 0.18 | 0.02 | 0.18 | 0.21 | 0.04 |
| | 100 | 0.00 | **0.22** | **0.03** | 0.08 | 0.09 | 0.11 |
| Grid 4 | 0.01 | 0.00 | **0.17** | 0.00 | 0.21 | **0.53** | 0.02 |
| | 0.1 | 0.00 | 0.16 | 0.01 | 0.29 | 0.44 | 0.02 |
| | 1 | 0.00 | 0.11 | 0.00 | **0.34** | 0.37 | 0.04 |
| | 10 | 0.11 | 0.10 | **0.02** | 0.05 | 0.33 | 0.02 |
| | 100 | **0.13** | **0.17** | 0.00 | 0.02 | 0.18 | **0.09** |
| Grid 5 | 0.01 | 0.00 | 0.13 | 0.01 | 0.26 | 0.49 | 0.03 |
| | 0.1 | 0.00 | **0.17** | 0.01 | 0.33 | 0.55 | 0.04 |
| | 1 | 0.00 | 0.14 | **0.06** | **0.44** | **0.56** | 0.06 |
| | 10 | 0.00 | 0.15 | 0.01 | 0.09 | 0.21 | 0.03 |
| | 100 | 0.00 | 0.11 | 0.02 | 0.01 | 0.14 | **0.16** |

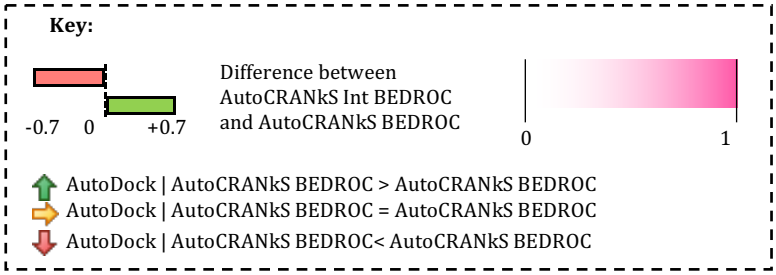



Figure 4.37 BEDROC (α = 80.5) values for rescoring of AutoCRANkS docked poses with AutoDock. The median, highest and lowest values across the five grids for each normalisation constant are shown. The highest BEDROC achieved for each grid is shown in bold. The arrows indicate whether the BEDROC calculated for AutoDock | AutoCRANkS is higher, the same as or lower than for AutoCRANkS. The bars indicate the difference in BEDROC between AutoDock | AutoCRANkS and AutoCRANkS. For HSP90A there is little change between AutoDock | AutoCRANkS and AutoCRANkS. For all other targets excluding BCL2 the average BEDROC is reduced for a normalisation constant of 1 on using AutoDock | AutoCRANkS.

For ADRB2, KIF11 and DHFR AutoCRANkS performed well with significant increases in AUC over AutoDock, in particular for a normalisation constant of 1. Figure 4.35 shows that for DHFR on average the AUC is significantly lower using AutoDock | AutoCRANkS. AutoCRANkS achieved an AUC of 0.79 for normalisation constants of 10 and 100 whereas AutoDock | AutoCRANkS achieved 0.48 and 0.46.

For ADRB2 and KIF11 the AUC calculated for AutoDock | AutoCRANkS is lower on average than for AutoCRANkS. Using a normalisation constant of 1 the AUC is reduced by -0.06 for ADRB2 and -0.30 for KIF11. For ADRB2 for all normalisation constants the AUC is also slightly lower than AutoDock. In contrast, for KIF11 the AUC AutoDock | AutoCRANkS is still significantly higher than for AutoDock This shows that for ADRB2 the binding poses generated by AutoCRANkS cause AutoDock to achieve a poorer discrimination between actives and decoys than when binding poses are generated by AutoDock. This indicates that the active grids may be detrimental to the pose generation. However, in the case of KIF11 the AUC is increased compared to AutoDock indicating that for this target the active grids could be causing the generation of binding poses that allow for better discrimination between actives and decoys.

The $EF_{0.5\%}$ for these targets calculated for docking by AutoDock | AutoCRANkS are on average worse than AutoCRANkS for all normalisation constants (Figure 4.36). The results for AutoCRANkS for these targets using a normalisation constant of 1, on average over the five grids, the $EF_{0.5\%}$ was higher than for GOLD, Glide, AutoDock Vina and AutoDock. However, when using AutoDock | AutoCRANkS for ADRB2, DHFR and KIF11 the $EF_{0.5\%}$ is no longer higher than for all the other docking tools

when using a normalisation constant of 1. However the $EF_{0.5\%}$ is still higher than for AutoDock. The BEDROC ($\alpha$ = 80.5) calculated for dockings by AutoDock | AutoCRANkS for targets DHFR, KIF11 and ADRB2 are on average lower than for AutoCRANkS (Figure 4.37). For KIF11 using a normalisation constant of 1, for AutoCRANkS on average the BEDROC ($\alpha$ = 80.5) was calculated to be 0.79, but for AutoDock | AutoCRANkS was calculated to be 0.34. For targets ADRB2 and DHFR on average the BEDROC ($\alpha$ = 80.5) calculated for results from AutoDock | AutoCRANkS for all normalisation constants was no better than for AutoDock.

For target BCL2 there is an increase in the AUC calculated for AutoDock | AutoCRANkS compared to AutoCRANkS (Figure 4.35). However, there is still a decrease in AUC as the normalisation constant increases in particular for normalisation constants of 10 (AUC = 0.59) and 100 (AUC = 0.53). None of the AUC values calculated for AutoDock | AutoCRANkS are higher than calculated for AutoDock. These results are mirrored by the $EF_{0.5\%}$ and BEDROC ($\alpha$ = 80.5) (Figure 4.36 and Figure 4.37 respectively). The values averaged across the five grids are higher than for AutoCRANkS but there is a large decrease for normalisation constants of 10 and 100, and the values averaged across the five grids are not greater than calculated for AutoDock. This indicates that as the normalisation constant increases the active grids are detrimental to both binding pose generation and scoring. Using the same binding poses AutoCRANkS achieved worse results for active versus decoy discrimination than when AutoDock was used to score the poses. Using the same scoring function AutoDock achieved similar or worse results when using AutoCRANkS generated binding poses as compared to using AutoDock generated binding poses.

For the target AKT1 the calculated AUCs for docked poses generated using AutoDock | AutoCRANkS are shown in Figure 4.35. The values on average for all normalisation constants are slightly worse than for AutoDock. However, for a normalisation constant of 10 for grids 1 (AUC = 0.71) and 3 (AUC = 0.69) the values are better than for AutoDock (AUC = 0.67) but not by a significant amount. There is a similar performance to AutoCRANkS for all normalisation constants on average. This could indicate that binding pose generation is causing a slight decrease in performance as when the poses generated by AutoCRANkS were scored by both AutoCRANkS and AutoDock the result was worse than when AutoDock poses were scored by AutoDock. The $EF_{0.5\%}$ for this target using AutoDock | AutoCRANkS show no significant change compared to both AutoDock and AutoCRANkS (Figure 4.36). This is mirrored by the BEDROC ($\alpha = 80.5$) where there is no significant difference between the values calculated for AutoDock | AutoCRANkS (for example BEDROC ($\alpha = 80.5$) = 0.07 for a normalisation constant of 1) and for AutoDock (BEDROC ($\alpha = 80.5$) = 0.01) as shown in Figure 4.37.

For grid 1 using a normalisation constant of 10 for the AKT1 dataset, there is an improvement in AUC, $EF_{0.5\%}$ and BEDROC ($\alpha = 80.5$) when using AutoDock | AutoCRANkS, compared to both AutoDock and AutoCRANkS. On investigation this improved performance is due to a low ranking of a few active compounds. One of these compounds (active compound 6 from the DEKOIS 2.0 AKT1 dataset) was ranked 213$^{th}$ when using AutoCRANkS but ranked 5$^{th}$ when the binding poses were rescored using AutoDock (AutoDock | AutoCRANkS). Each of the ten binding poses for this active compound are shown in Figure 4.38 and Figure 4.39. The poses are coloured by the per atom contribution to the score.

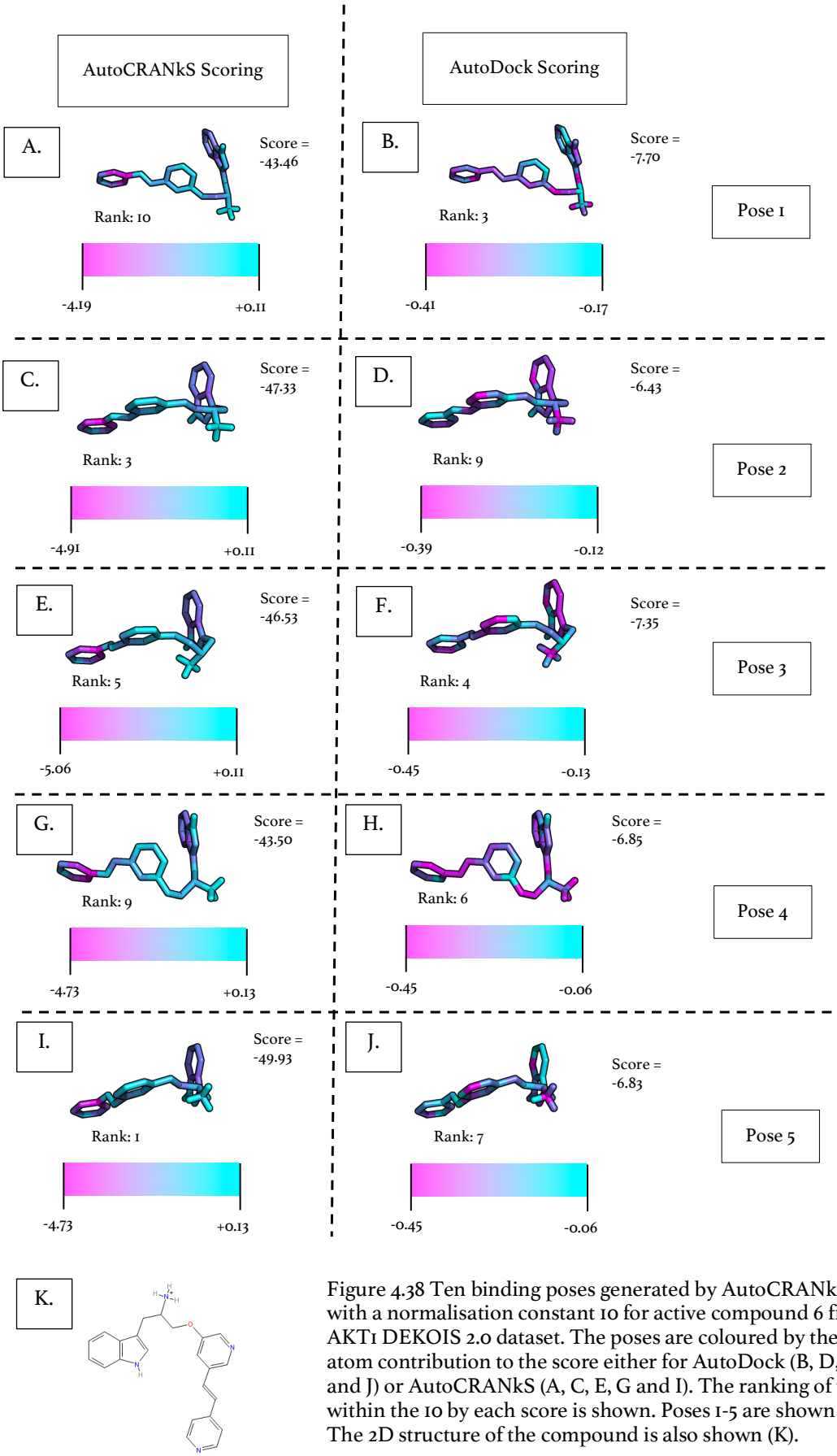



Figure 4.38 Ten binding poses generated by AutoCRANkS grid 1 with a normalisation constant 10 for active compound 6 from the AKT1 DEKOIS 2.0 dataset. The poses are coloured by the per atom contribution to the score either for AutoDock (B, D, F, H, and J) or AutoCRANkS (A, C, E, G and I). The ranking of the pose within the 10 by each score is shown. Poses 1-5 are shown here. The 2D structure of the compound is also shown (K).

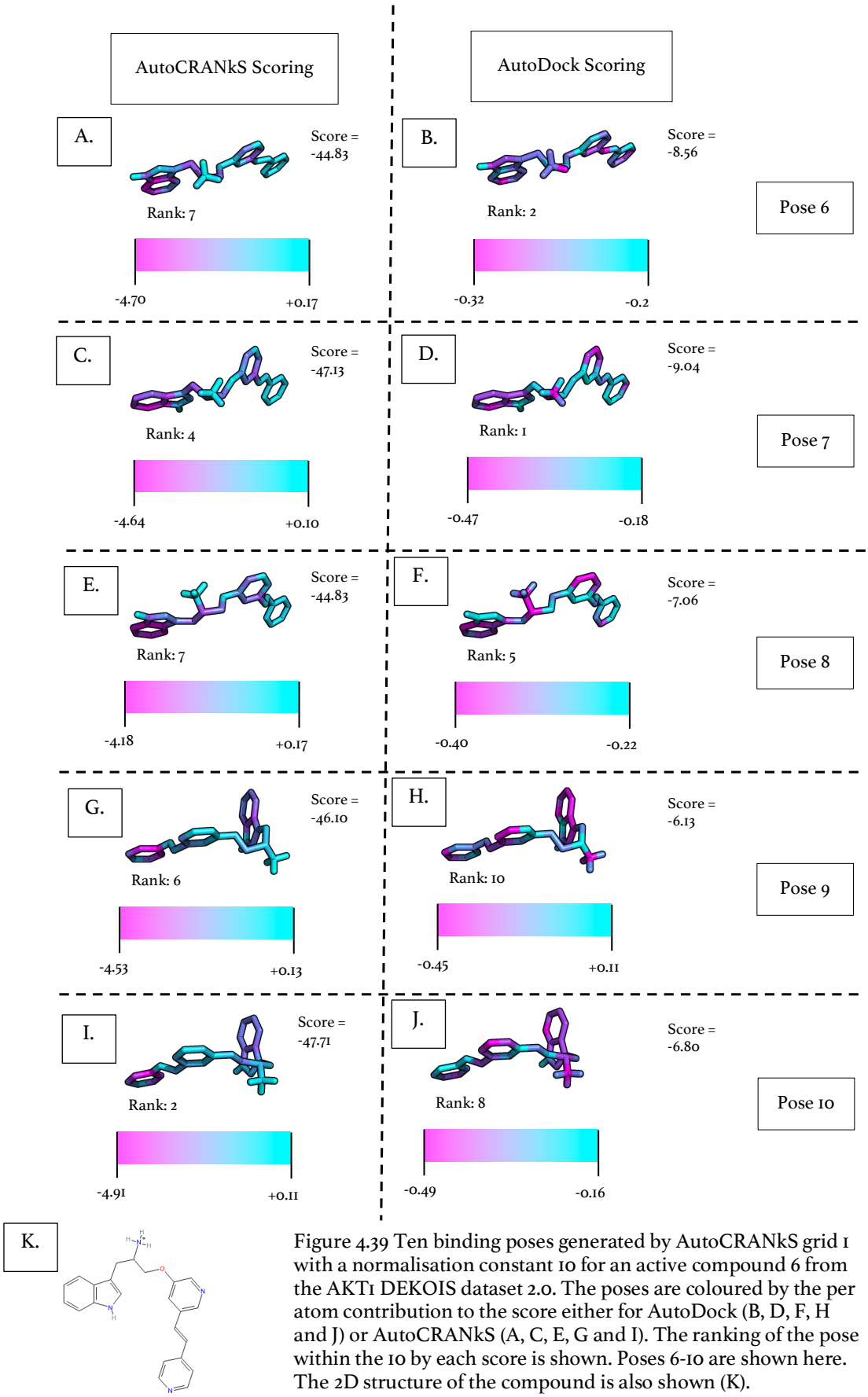



Figure 4.39 Ten binding poses generated by AutoCRANkS grid 1 with a normalisation constant 10 for an active compound 6 from the AKT1 DEKOIS dataset 2.0. The poses are coloured by the per atom contribution to the score either for AutoDock (B, D, F, H and J) or AutoCRANkS (A, C, E, G and I). The ranking of the pose within the 10 by each score is shown. Poses 6-10 are shown here. The 2D structure of the compound is also shown (K).

AutoDock scored pose 7 significantly lower than the rest of the poses (Figure 4.39 C). AutoCRANkS also scored this pose low (ranked 4$^{th}$ out of the ten poses) but not significantly lower than the other poses. The score for the pose calculated by AutoCRANkS is dominated by the link between two aromatic rings (Figure 4.39 C). The score is dominated by aromatic contributions and some carbon contributions. In contrast for the AutoDock scoring all the nitrogen atoms of the compound are making a significant contribution to the score, as well as some aromatic carbons and oxygens (Figure 4.39 D).

Poses 6 and 8 are calculated to have the same score by AutoCRANkS (Figure 4.39 A and E) but not by AutoDock (Figure 4.39 B and F). The scoring pattern per atom for AutoCRANkS is very similar between the two poses. Atoms, although in slightly different positions are overlapping with the same hot spots of potential. Again, contributions are from aliphatic and aromatic carbon atoms, not heteroatoms (Figure 4.39 A and E). AutoDock however calculates very different scoring patterns for the poses (Figure 4.39 B and F). For pose 6 some aromatic carbon and nitrogen atoms, some oxygen atoms and some hydrogen atoms are contributing to the score. For pose 7 there are big contributions from an aromatic nitrogen atom and the nitrogen of the primary amine to the score, as well as some hydrogen, oxygen and carbon atoms. This indicates that the larger wells of potential caused by the overlap of the Gaussian functions of atoms in the active grids may not be specific enough, and that the finer level of detail that the original maps allow scoring for this target by AutoDock to be more beneficial.

This indicates that the contribution of heteroatoms to the score could be crucial for the performance of docking tools for this target and dataset. This could also explain why for this target there was an improvement when using AutoCRANkS Int over AutoCRANkS or AutoDock. For AutoCRANkS the potential is dominated by aromatic and carbon contributions to the binder grids without much contribution from the heteroatoms to the score. However, for AutoCRANkS Int the interaction filtering removes a large proportion of the aromatic and carbon atom from the binder grids. This is shown in Table 4.6. This could allow the oxygen and nitrogen maps to contribute more greatly to the scoring of AutoCRANkS Int causing the increase in performance of this method.

| AutoDock Atom Type | AutoCRANkS | AutoCRANkS Int |
|---|---|---|
| Aromatic Carbon | 30 | 2 |
| Aliphatic Carbon | 37 | 2 |
| Nitrogen Acceptor | 9 | 3 |
| Oxygen Acceptor | 25 | 13 |

Table 4.6 The number of atoms in the active grids for grid 1 for target AKT1 that will be added to the AutoGrid map when using AutoCRANkS or AutoCRANkS Int.

### *4.3.3.1 Summary of AutoDock | AutoCRANkS*

To determine the effect of active grids on the generation of binding poses, the poses generated by AutoCRANkS were rescored using AutoDock. The results were found to be target-dependent. For targets for which AutoCRANkS achieved significantly higher AUC, $EF_{0.5\%}$ and BEDROC ($\alpha$ = 80.5) values than AutoDock (ADRB2, DHFR

and KIF11) the values calculated for AutoDock | AutoCRANkS were lower than for AutoCRANkS. For KIF11 however the values were still higher than for AutoDock, indicating that the poses generated by AutoCRANkS allow for better discrimination between actives and decoys for this target. For ADRB2 and DHFR there is little difference between the values calculated for AutoDock | AutoCRANkS and AutoDock.

For BCL2 the detriment to the AUC values calculated for AutoDock when the protocol is changed to AutoCRANkS is also exhibited for AutoDock | AutoCRANkS. This indicates that the active grids are having a detrimental effect on the binding pose generation for this target.

For target AKT1 on average there was little difference between the performance of AutoDock | AutoCRANkS and AutoDock or AutoCRANkS. However, for grids 1 and 3 there was increased AUC, $EF_{0.5\%}$ and BEDROC ($\alpha$ = 80.5) values for a normalisation constant of 10 compared to AutoDock and AutoCRANkS. This could be attributed to the fact the AutoDock scoring pattern of the binding poses allows more significant contributions from heteroatoms, whereas the AutoCRANkS scoring pattern is dominated by contributions from aromatic and carbon atoms.

Overall only one target showed an improvement in results when using AutoDock | AutoCRANkS over AutoCRANkS and consequently this method should not be considered.

### *4.3.4 Scoring of AutoDock Poses by AutoCRANkS*

To better determine the separate effects of AutoCRANkS on the binding pose generation and scoring function, the docked poses generated by AutoDock were rescored using the AutoCRANkS scoring function for each of the targets. This will be referred to as AutoCRANkS | AutoDock. Five different grids with five normalisation constants were used for each target. The ability to discriminate actives from decoys is measured by calculating the AUC (Figure 4.40), $EF_{0.5\%}$ (Figure 4.41) and the BEDROC ($\alpha$ = 80.5) (Figure 4.42).

For target HSP90A on average across the five grids the AUC calculated for AutoCRANkS | AutoDock is slightly higher than for AutoCRANkS for all normalisation constants (a change of between 0.02 and 0.12). However, the best results across any of the grids were higher for AutoCRANkS than for AutoCRANkS | AutoDock. For a normalisation constant of 1 the highest AUC achieved using grid 3, for AutoCRANkS was 0.71 whereas for AutoCRANkS | AutoDock the highest AUC was 0.67. However, on average the AUCs remain higher than for AutoDock indicating an improvement of scoring function when using AutoCRANkS over AutoDock. There is again little change for $EF_{0.5\%}$ with on average zero enrichment for all normalisation constants calculated for AutoCRANkS | AutoDock (Figure 4.41). There is also very little difference in the BEDROC ($\alpha$ = 80.5) between AutoCRANkS and AutoCRANkS | AutoDock for this target (Figure 4.42).

| Target | Normalisation Constant | HSP90A | ADRB2 | DHFR | KIF11 | BCL2 | AKT1 |
|---|---|---|---|---|---|---|---|
| Median | 0.01 | 0.41 | 0.65 | 0.43 | 0.78 | **0.88** | 0.63 |
| | 0.1 | 0.49 | **0.68** | 0.61 | **0.82** | **0.88** | **0.65** |
| | 1 | 0.64 | 0.63 | **0.72** | 0.81 | 0.79 | **0.65** |
| | 10 | **0.65** | 0.64 | **0.72** | 0.80 | 0.75 | 0.63 |
| | 100 | **0.65** | 0.63 | **0.72** | 0.80 | 0.75 | 0.62 |
| Best | 0.01 | 0.42 | 0.65 | 0.43 | 0.78 | **0.88** | 0.63 |
| | 0.1 | 0.51 | **0.72** | 0.63 | **0.82** | **0.88** | 0.66 |
| | 1 | 0.67 | 0.68 | **0.73** | 0.81 | 0.80 | **0.67** |
| | 10 | 0.69 | 0.68 | **0.73** | 0.81 | 0.76 | 0.65 |
| | 100 | **0.70** | 0.68 | **0.73** | 0.80 | 0.76 | 0.65 |
| Worst | 0.01 | 0.41 | 0.63 | 0.43 | 0.78 | **0.88** | 0.62 |
| | 0.1 | 0.42 | **0.67** | 0.57 | **0.81** | 0.87 | **0.64** |
| | 1 | 0.49 | 0.59 | **0.71** | 0.78 | 0.76 | 0.59 |
| | 10 | **0.51** | 0.59 | **0.71** | 0.75 | 0.70 | 0.57 |
| | 100 | **0.51** | 0.59 | **0.71** | 0.75 | 0.70 | 0.56 |
| Grid 1 | 0.01 | 0.41 | 0.65 | 0.43 | 0.78 | **0.88** | 0.62 |
| | 0.1 | 0.42 | **0.68** | 0.61 | **0.82** | **0.88** | **0.65** |
| | 1 | 0.49 | 0.62 | **0.72** | 0.81 | 0.8 | 0.64 |
| | 10 | **0.51** | 0.61 | **0.72** | 0.8 | 0.75 | 0.61 |
| | 100 | **0.51** | 0.61 | **0.72** | 0.8 | 0.75 | 0.61 |
| Grid 2 | 0.01 | 0.42 | 0.63 | 0.43 | 0.78 | **0.88** | 0.63 |
| | 0.1 | 0.43 | **0.68** | 0.6 | **0.81** | **0.88** | **0.65** |
| | 1 | **0.52** | 0.63 | **0.71** | 0.78 | 0.8 | **0.65** |
| | 10 | **0.52** | 0.64 | **0.71** | 0.75 | 0.76 | 0.63 |
| | 100 | 0.51 | 0.63 | **0.71** | 0.75 | 0.76 | 0.63 |
| Grid 3 | 0.01 | 0.41 | 0.65 | 0.43 | 0.78 | **0.88** | 0.62 |
| | 0.1 | 0.51 | **0.7** | 0.63 | **0.81** | **0.88** | **0.64** |
| | 1 | 0.67 | 0.68 | 0.72 | 0.78 | 0.78 | 0.59 |
| | 10 | 0.69 | 0.68 | **0.73** | 0.77 | 0.73 | 0.57 |
| | 100 | **0.7** | 0.68 | **0.73** | 0.77 | 0.73 | 0.56 |
| Grid 4 | 0.01 | 0.41 | 0.65 | 0.43 | 0.78 | **0.88** | 0.63 |
| | 0.1 | 0.49 | **0.67** | 0.57 | **0.82** | 0.87 | 0.66 |
| | 1 | 0.64 | 0.59 | 0.71 | 0.81 | 0.79 | **0.67** |
| | 10 | **0.65** | 0.59 | **0.72** | 0.81 | 0.75 | 0.65 |
| | 100 | **0.65** | 0.59 | **0.72** | 0.8 | 0.75 | 0.65 |
| Grid 5 | 0.01 | 0.41 | 0.63 | 0.43 | 0.78 | **0.88** | 0.63 |
| | 0.1 | 0.51 | **0.72** | 0.61 | **0.82** | 0.87 | **0.66** |
| | 1 | 0.64 | 0.68 | **0.73** | 0.81 | 0.76 | 0.65 |
| | 10 | **0.66** | 0.68 | **0.73** | 0.8 | 0.7 | 0.63 |
| | 100 | **0.66** | 0.68 | **0.73** | 0.8 | 0.7 | 0.62 |

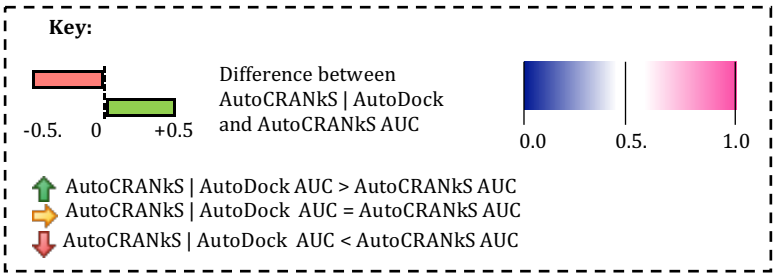



Figure 4.40 AUC values for docking of the targets from the DEKOIS 2.0 dataset using AutoDock poses rescored by AutoCRANkS. The median, highest and lowest values across the five grids for each normalisation constant are shown. The difference between the AUC for AutoCRANkS | AutoDock and AutoCRANkS is indicated by the data bars and arrows.

| Target | Normalisation Constant | HSP90A | ADRB2 | DHFR | KIF11 | BCL2 | AKT1 |
|---|---|---|---|---|---|---|---|
| Median | 0.01 | 0 | 0 | 0 | 5 | **20** | 0 |
| | 0.1 | 0 | 5 | 1 | 15 | 16 | 0 |
| | 1 | 0 | 10 | **15** | **31** | **20** | 5 |
| | 10 | 0 | **11** | **15** | **31** | **20** | **10** |
| | 100 | 0 | **11** | **15** | **31** | 16 | **10** |
| Best | 0.01 | 0 | 0 | 0 | 5 | 20 | 0 |
| | 0.1 | 0 | 5 | 5 | 21 | **25** | 0 |
| | 1 | 0 | 11 | **21** | **31** | **26** | **15** |
| | 10 | 1 | **15** | **21** | **31** | 21 | **15** |
| | 100 | 3 | **15** | **21** | **31** | 21 | **15** |
| Worst | 0.01 | 0 | 0 | 0 | 5 | **20** | 0 |
| | 0.1 | 0 | 5 | 0 | **15** | 16 | 0 |
| | 1 | 0 | 1 | **10** | 5 | **20** | 5 |
| | 10 | 0 | **10** | **10** | 5 | 10 | 5 |
| | 100 | 0 | 6 | 5 | 5 | 10 | 5 |
| Grid 1 | 0.01 | 0 | 0 | 0 | 5 | **20** | 0 |
| | 0.1 | 0 | 5 | 0 | 15 | 16 | 0 |
| | 1 | 0 | **10** | **10** | **31** | **20** | **15** |
| | 10 | 0 | **10** | **10** | **31** | 10 | **15** |
| | 100 | 0 | **10** | 5 | **31** | 10 | **15** |
| Grid 2 | 0.01 | 0 | 0 | 0 | 5 | 20 | 0 |
| | 0.1 | 0 | 5 | 5 | **15** | 16 | 0 |
| | 1 | 0 | 10 | **15** | 5 | **26** | **5** |
| | 10 | 0 | **11** | **15** | 5 | 21 | **5** |
| | 100 | 0 | 6 | **15** | 5 | 21 | **5** |
| Grid 3 | 0.01 | 0 | 0 | 0 | 5 | 20 | 0 |
| | 0.1 | 0 | 5 | 1 | **16** | **25** | 0 |
| | 1 | 0 | 1 | **21** | **16** | 21 | **10** |
| | 10 | 0 | **11** | 20 | 15 | 20 | **10** |
| | 100 | 0 | **11** | 20 | 15 | 16 | **10** |
| Grid 4 | 0.01 | 0 | 0 | 0 | 5 | **20** | 0 |
| | 0.1 | 0 | 5 | 5 | 21 | 16 | 0 |
| | 1 | 0 | 11 | **11** | **31** | **20** | **5** |
| | 10 | 1 | **15** | 10 | **31** | **20** | **5** |
| | 100 | 3 | **15** | 6 | **31** | 16 | **5** |
| Grid 5 | 0.01 | 0 | 0 | 0 | 5 | **20** | 0 |
| | 0.1 | 0 | 5 | 1 | 15 | 16 | 0 |
| | 1 | 0 | 5 | 16 | **31** | **20** | 5 |
| | 10 | 0 | **11** | **21** | **31** | 15 | **11** |
| | 100 | 0 | **11** | **21** | **31** | 15 | **11** |

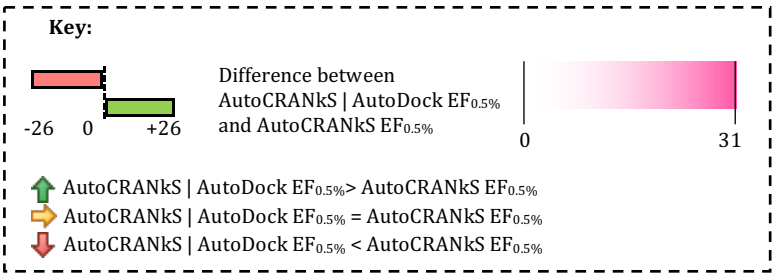



Figure 4.41 $EF_{0.5\%}$ values for docking of the targets from the DEKOIS 2.0 dataset using AutoDock poses rescored by AutoCRANkS. The median, highest and lowest values across the five grids for each normalisation constant are shown. Highest and lowest values across the five grids for each normalisation constant are shown. The difference between the enrichment calculated for AutoCRANkS | AutoDock and AutoCRANkS is indicated by the data bars and arrows.

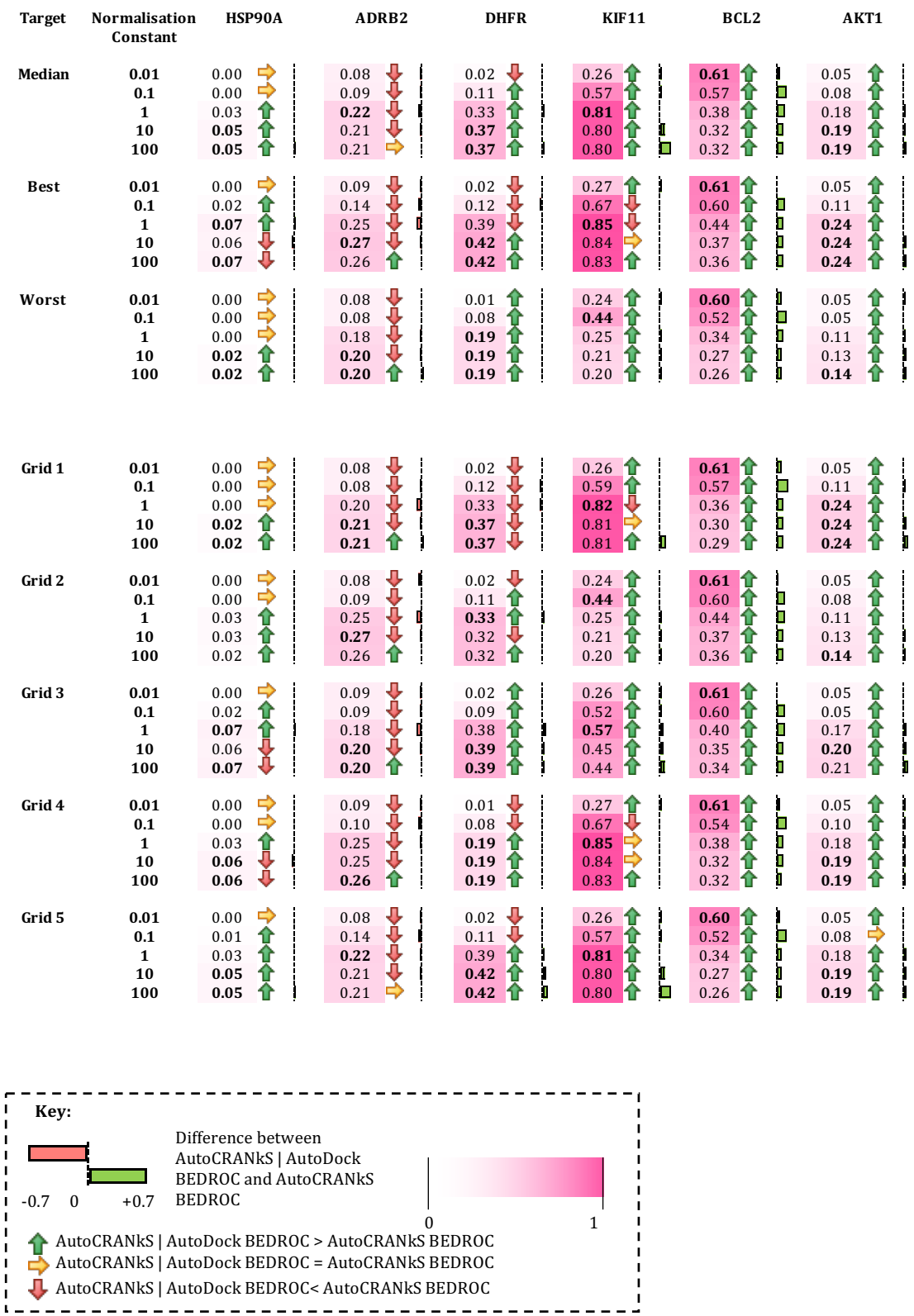



Figure 4.42 BEDROC (α = 80.5) values for docking of the targets from the DEKOIS 2.0 dataset using AutoDock poses rescored by AutoCRANkS. The median highest and lowest values across the five grids for each normalisation constant are shown. The difference between the BEDROC for AutoCRANkS | AutoDock and AutoCRANkS are shown by arrows and data bars.

This is indicative that the active grids improve the scoring function for this target compared to the AutoDock scoring function, as the AUC when the AutoDock poses are scored by AutoCRANkS is higher than when AutoDock scores the poses. The lack of enrichment measured by $EF_{0.5\%}$ and BEDROC ($\alpha$ = 80.5) again indicates that the improved AUC is due to the down-ranking of decoys rather than the up-ranking of active compounds.

For targets ADRB2, DHFR and KIF11 AutoCRANkS showed high performance with significantly higher AUC, $EF_{0.5\%}$ and BEDROC ($\alpha$ = 80.5) than for AutoDock. For AutoCRANkS | AutoDock for these targets on average the AUCs calculated on average are not as high as calculated for AutoCRANkS. The reduction ranges between 0.01 and 0.13 for these targets when averaged across the five grids. For ADRB2 an AUC of 0.74 was calculated for a normalisation constant of 0.1 on average for AutoCRANkS, whereas the corresponding AUC for AutoCRANkS | AutoDock is 0.68. For KIF11 an average AUC of 0.94 using AutoCRANkS with a normalisation constant of 1 is reduced to 0.81. For DHFR an average AUC of 0.78 for a normalisation constant of 1 when using AutoCRANkS is lowered to 0.72 for AutoCRANkS | AutoDock. However, for targets DHFR and KIF11 the average AUC at normalisation constants of 1 and above is still significantly higher than the AUC calculated for AutoDock. This indicates that for DHFR and KIF11 the addition of the active grids to AutoCRANkS causes an increase in performance of the scoring function compared to AutoDock. However, for ADRB2 the AUC calculated for AutoCRANkS | AutoDock is not significantly different from the AUC calculated for

AutoDock indicating that for this target there is not the same improvement to scoring function.

The $EF_{0.5\%}$ calculated for these three targets is shown in Figure 4.41. For ADRB2 and DHFR there is little difference to the average $EF_{0.5\%}$ calculated for AutoCRANkS | AutoDock compared to the $EF_{0.5\%}$ calculated for AutoCRANkS. However, for KIF11 for normalisation constants of 10 and 100 there is a large increase in $EF_{0.5\%}$ compared to AutoCRANkS (5 and 26 respectively). This indicates that for KIF11 as the normalisation constant increases above 1, the decreased performance is due to the binding pose generation being affected by domination of the active grids, as the AutoCRANkS scoring function is able to perform equally well for normalisation constants of 10 and 100 for AutoCRANkS | AutoDock.

This behaviour is mirrored by calculated BEDROC ($\alpha$ = 80.5) where there is little difference between the calculated values for AutoCRANkS and for AutoCRANkS | AutoDock for a normalisation constant of 1 for ADRB2, DHFR and KIF11. The only significant difference is again the improvement on average for KIF11 for normalisation constants of 10 and 100 calculated for AutoCRANkS | AutoDock compared to the corresponding values calculated for AutoCRANkS.

For target AKT1 there is no significant change in calculated values for AutoCRANkS | AutoDock compared to AutoCRANkS for either the AUC, $EF_{0.5\%}$ or BEDROC ($\alpha$ = 80.5). For target BCL2 however, AutoCRANkS | AutoDock exhibits an improvement in performance over AutoCRANkS. Using a normalisation constant of 1 the AUC calculated for AutoCRANkS | AutoDock is 0.79 – a 0.37 increase compared to the

AUC calculated from results generated by AutoCRANkS. Although this is still worse than for AutoDock, the reduction in AUC is much less (an AUC of 0.85 was achieved for AutoDock).

### *4.3.4.1 Summary of AutoCRANkS | AutoDock*

The binding poses generated by AutoCRANkS were rescored using the AutoDock scoring function. For targets ADRB2 and DHFR there is an improvement on average over AutoDock but values of AUC, $EF_{0.5\%}$ and BEDROC ($\alpha$ = 80.5) are lower than for AutoCRANkS. For KIF11 AUC values are on average reduced compared to AutoCRANkS but for $EF_{0.5\%}$ and BEDROC ($\alpha$ = 80.5) there is an increase in performance of AutoCRANkS | AutoDock compared to AutoCRANkS. For target BCL2 the detriment to values as the normalisation constant increases is still exhibited for AutoCRANkS | AutoDock. However, the reduction in values is significantly less. For target HSP90A there is an increase in AUC on average compared to AutoDock and AutoCRANkS, For AKT1 there is little change in AUC compared to AutoCRANkS. However, there is an increase in enrichment compared to AutoCRANkS.

Overall the AutoCRANkS | AutoDock protocol shows potential as a docking protocol. For a normalisation constant of 1 the AUC for HSP90A is increased from 0.57 to 0.64, the AUC for BCL2 is increased from 0.42 to 0.79 and for AKT1 from 0.62 to 0.65 on average on moving from AutoCRANkS to AutoCRANkS | AutoDock. However, the AUC for ADRB2 is reduced from 0.68 to 0.63, for DHFR from 0.78 to 0.72 and for KIF11 from 0.94 to 0.81. For a normalisation constant of 1 AutoCRANkS |

AutoDock either improves or keeps the same enrichment for all targets excluding ADRB2. For AutoCRANkS | AutoDock the enrichment on average for a normalisation constant of 1 is higher or the same as the enrichment for 3 out of the 4 docking tools compared to here for all the targets. For three of the targets the enrichment is higher than for all of the docking tools. For the BEDROC ($\alpha$ = 80.5) the average value for AutoCRANkS | AutoDock using a normalisation constant of 1 is higher than for AutoCRANkS and either higher or similar to the calculated BEDROC ($\alpha$ = 80.5) for AutoDock. AutoCRANkS | AutoDock gives more stable results than for AutoCRANkS: whilst the best results may not be as good the worst results are not as bad, and there is an increase in enrichment.

## 4.4 Conclusions

We have developed several active-guided docking protocols that incorporate knowledge about active ligands from protein-ligand structures into docking by AutoDock. This was achieved by modifying the AutoGrid maps that describe atomic affinities and mapping the corresponding atom types from CRANkS maps. By making use of the structural data that may be available, for example from fragment screening, the docking protocol can be modified to bias the binding pose generation and scoring function of the target of interest.

The AutoCRANkS protocol involves adding all atoms from a set of protein-aligned bound ligands to the AutoGrid maps for a target. This was tested on six targets from the DEKOIS 2.0 dataset using five different grids assembled from ligands complexed

with the same target. Five different weights of the Gaussian functions were investigated.

First for the remaining protein-ligand structures not used in a given AutoCRANkS grid the ligands were redocked using AutoDock and AutoCRANkS protocols. The results were target-dependent but for targets with a large number of structures there was improvement in the RMSD between the X-ray crystal structure and the lowest scored conformer and for the average across all conformers when using AutoCRANkS over AutoDock. However, there was no consensus across targets for the optimum normalisation value.

Secondly the actives and decoys of each of the six targets from the DEKOIS 2.0 dataset were docked using AutoCRANkS and AutoDock. For four out of six targets there was a significant improvement on average for the AUC when using AutoCRANkS compared to AutoDock, in particular for a normalisation constant of 1. The $EF_{0.5\%}$ on average using a normalisation constant of 1 was higher when using AutoCRANkS than for AutoDock Vina, AutoDock, GOLD and Glide. Of the remaining three targets only one target was found to have significantly worse enrichment than AutoDock for a normalisation constant of 1. These results indicate that AutoCRANkS is an improved docking protocol over AutoDock and that a normalisation constant of 1 should be used. I have shown that by using structural knowledge about bound ligands it is possible to improve our ability to discriminate between actives and decoys,

A variant of AutoCRANkS, AutoCRANkS Int, was also developed and validated on the actives and decoys from DEKOIS 2.0 for each target. This protocol filters the ligand atoms used in the active grids by only including atoms that are calculated to form interactions. For targets that were calculated to have a high AUC, $EF_{0.5\%}$ and BEDROC ($\alpha$ = 80.5) values when docked by AutoCRANkS, these values were reduced by AutoCRANkS Int. For BCL2 which gave extremely low values when docked by AutoCRANkS, the values for AutoCRANkS Int are higher as the contribution of the active grids is reduced and the number of atoms is reduced. For AKT1 there is an improvement on using AutoCRANkS Int over AutoCRANkS, indicating that for this target the filtering of non-interacting atoms is reducing the noise. However, overall AutoCRANkS Int was not found to reduce noise and increase valuable signal from the data. In general, the ability to discriminate between actives and decoys was reduced with the elimination of non-interacting atoms from the active grids.

To further explore the performance of the protocols, the AutoCRANkS generated poses were rescored using AutoDock (AutoDock | AutoCRANkS) and the AutoDock generated poses were rescored using AutoCRANkS (AutoCRANkS | AutoDock). For AutoCRANkS | AutoDock, on average using a normalisation constant of 1 there was an increase in AUC over AutoCRANkS for three out of six targets and a reduction for the other three. However, there was an increase in enrichment on average using a normalisation constant of 1 for five out of six targets. For AutoCRANkS | AutoDock the $EF_{0.5\%}$ on average using a normalisation constant of 1 is higher than for three out of the four docking tools: AutoDock, AutoDock Vina, GOLD and Glide. Overall AutoCRANkS | AutoDock gives more stable results than AutoCRANkS: although the best results can be worse, the worst results are better.

To better determine the appropriate docking protocol to use out of AutoCRANkS and AutoCRANkS | AutoDock, both protocols as well as AutoDock should be run on the remaining 75 targets from the DEKOIS 2.0 dataset. This will determine the performance of the protocols across a wide variety of targets. The protocols could also be run on targets from the DUD-E dataset as this will allow comparison to a high number of other tools as shown in Chapter 3.

Additionally, the place of the docking protocol in the pipeline for the selection of compounds should be explored. In Chapter 3 I described a multi-objective optimisation algorithm to select compounds for follow-up. It is well-known that the potency-trap should be avoided, and compounds should not solely be selected on docking rankings. Further work should explore combining the docking scores from this Chapter with multi-objective optimisation and the effect the protocols can have on the selection of compounds.

# Chapter 5

# Case Studies

## 5.1 Introduction

The methodologies described so far in this thesis were developed to aid the selection of follow-up compounds from initial hits for a given protein target. In this Chapter I describe the use of CRANkS as part of prospective studies at Diamond Light Source. The XChem group at Diamond Light Source use fragment screening to identify hits for protein targets using X-ray crystallography (*http://www.diamond.ac.uk/Beamlines/Mx/Fragment-Screening.html*). The whole screen is now a highly streamlined procedure, allowing fast collection of results. The CRANkS algorithm was developed to make use of this data to select follow-up compounds. The XChem group routinely investigates targets of interest being explored by the SGC Oxford. This testing allows CRANkS to be used in a prospective setting to identify any issues that may be apparent when using the algorithm in a "real-world" scenario.

One protein family explored by the SGC Oxford is the NUDIX family. This family of of 24 proteins cleaves nucleoside diphosphates that are linked to any moiety (McLennan, 2006). The proteins control levels of metabolic intermediates and signalling compounds in various parts of the body (Bessman *et al.,* 1996). The most studied NUDIX protein is MTH1, an enzyme crucial to the process of DNA repair. The over-expression of the protein in mice models was found to increase their

longevity linking aging to the oxidative damage of nucleic acids (De Luca *et al.,* 2013). The overexpression of MTH1 in tumour cells led its inhibition to be used as cancer treatment, however studies produced conflicting results as to the effect of MTH1 inhibitors on the cytotoxicity of cancer cells (Samaranayake *et al.,* 2017). There is some literature about NUDT21 (Brumbaugh *et al.,* 2018), DCP2 (Li *et al.,* 2012) and NUDT5, with specific inhibitors determined for NUDT5 (Page *et al.*, 2018), but the SGC successfully determined structures for many more proteins in the family, on which there has been little study.

The SGC made the NUDIX family a strategic priority three years ago to increase the structural coverage with a view of enabling chemical probe development for these proteins. This should enable others to study them further. Fragment-screening by the XChem group is one of the methods the SGC use to identify small molecule starting points for chemical probe development. I used CRANkS to select follow-up compounds from the results of fragment screening for two members of the NUDIX family: NUDT22 and NUDT7.

The work for NUDT22 was carried out during the work presented in Chapters 2 and 3 and the work for NUDT7 was carried out during the work presented in Chapter 3. The compounds I proposed for follow-up for both targets have not yet been purchased and synthesised. For target NUDT22 the target was deprioritised and for target NUDT7 a different strategy was employed.

For NUDT22 two subsets were proposed for synthesis based on two hypotheses. The first is that by minimising the Interaction Score compounds will be selected that

form the same interactions as the fragment hits and are more likely to be active. The second is that by maximising the Interaction Score compounds will be selected that form novel interactions and so can be synthesised to explore new areas of interaction space. For NUDT7 compounds were selected with a high Interaction Score to facilitate synthesis of compounds that form novel interactions and explore the binding pocket.

Working on both targets highlighted some issues with the CRANkS algorithm. The first is the conformer generation step which currently does not have any threshold for how different conformers generated must be to each other i.e. 100 identical conformers can be generated. In Chapter 4 this issue is circumvented by making use of the AutoDock conformer generation tool which uses a genetic algorithm and subsequent conformational cluster analysis. Secondly, there was no robust protocol to make use of the scores to select compounds – this was developed in Chapter 3 and the multi-objective optimisation method was shown to prioritise novel active scaffolds. Thus, this is the procedure that should be used to select compounds in the future and the improved methodology in Chapter 3 will be installed at Diamond Light Source to allow the algorithms to be used in the future. With the further exploration and much more rigorous testing of the method in Chapter 3, a more convincing argument can be made to use the CRANkS algorithm, than I was able to when completing the work for this chapter.

## 5.2 NUDT22

### *5.2.1 Introduction*

Target NUDT22 is a member of the NUDIX protein family whose structure was determined by X-ray crystallography by the SGC Oxford (PDB Code 5lf9; UniProt Code Q9BRQ3; C.Tallant *et al.,* 2017). The target was screened by the XChem group against a poised fragment library (Cox *et al.,* 2016). This is a library designed to be amenable to follow-up using one step chemistry by the top ten most commonly used reactions in drug discovery (Roughley and Jordan, 2011) and a further 12 heterocycle forming reactions (Hartenfeller *et al*., 2011).

The library was soaked against crystals of NUDT22 and structures determined using X-ray crystallography. Hits are fragments that were found to bind to the target using this method. High confidence hits are defined as ligands that were "easily interpretable from clear density and refinement was well-behaved" (http://thesgc.org/ligand-bounds/nudt22-1) i.e. the structure of the ligand was modelled with confidence by the crystallographer and so can be trusted. For target NUDT22 16 high-confidence hits were found and are discussed in Section 5.2.2. These hits were used to generate a follow-up library of compounds and the methods in which the CRANkS algorithm could be used to select compounds for synthesis from this library were investigated.

The general procedure for selecting follow-up compounds is shown in Figure 5.1. Follow-up compounds are generated using XPoise developed by Oakley Cox (Cox, 2015) (described below in Section 5.2.2) using the fragment hits for the target. The hits are also used to construct CRANkS grids. The follow-up compounds are then scored

using the CRANkS grids, generating Element, Pharmacophore and Interaction Scores. Using these scores compounds can then be selected for synthesis. Several methods of selection are discussed for this target.

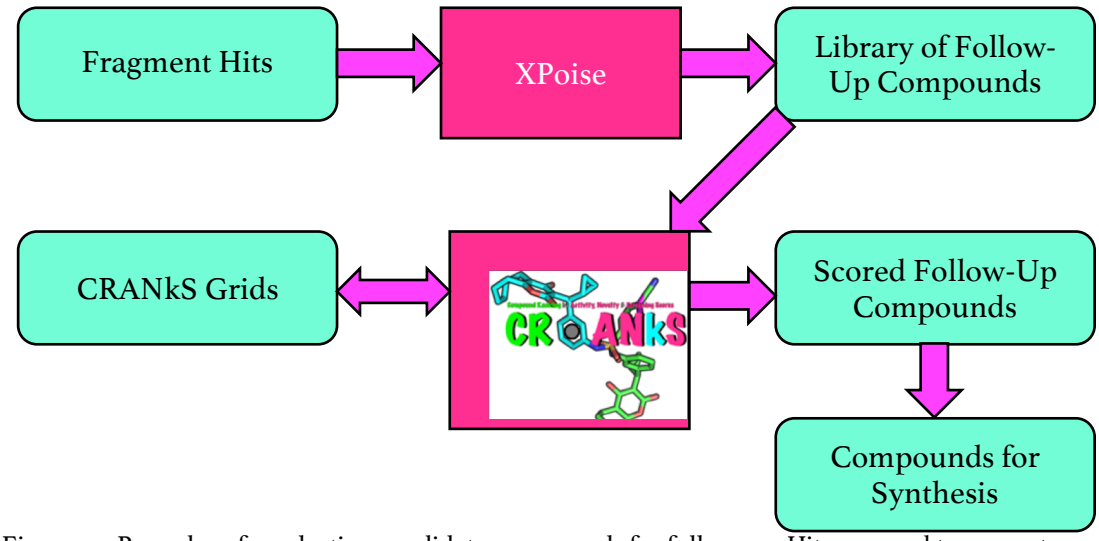



Figure 5.1. Procedure for selecting candidate compounds for follow-up. Hits are used to generate a library of potential follow-ups from the poised fragments using the KNIME workflow XPoise. CRANkS grids are also constructed from the hits and follow-up compounds are scored using the grids. Compounds can then be selected for synthesis based on these scores. Various methods of selection are discussed in this Chapter.

## *5.2.2 Method*

All fragment hits for NUDT22 determined by the SGC can be found on the SGC website (https://www.thesgc.org/ligand-bounds/nudt22-1). Each of the file names from the website and the corresponding abbreviation is shown in Appendix B, Table B.5.1, with the SMILES for each of the ligands. The corresponding structure for each of the hits are shown in Figure 5.2. and the collated structures are shown in Figure 5.3.

Figure 5.3 shows that within the main binding site most ligands adopt a conserved shape and location with high overlap and are found in site 1. Two hits are within the main binding site but are bound in another area of the binding site  shown in Figure 5.3 as site 2. These are structures x0243 and one of the ligand copies from structure x0826. One high confidence hit (structure x0290) was found to bound in a different binding site and was subsequently not used to construct any CRANkS grid. Each of the models was visually inspected and were pre-aligned by the crystallographer.

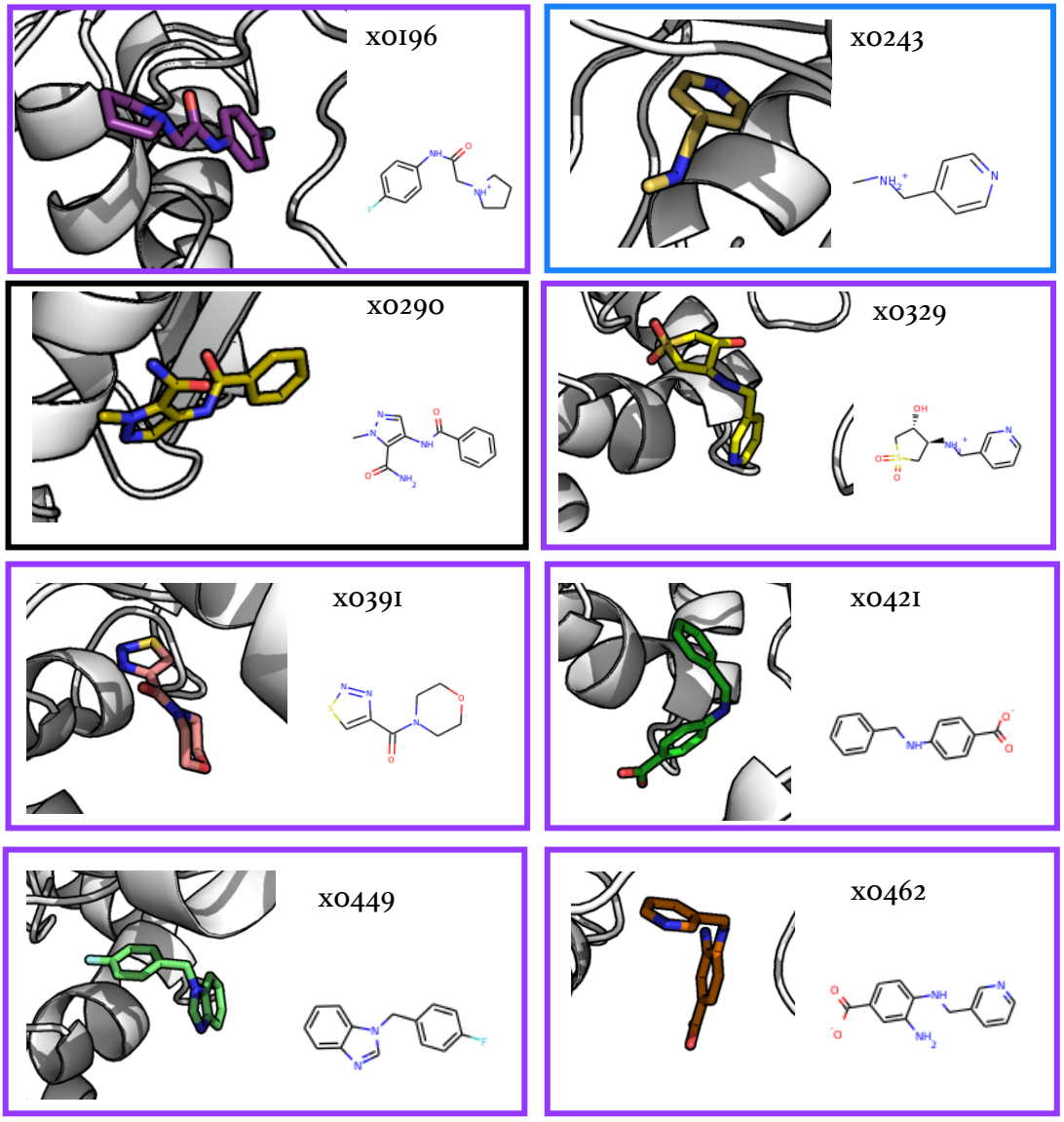


Figure 5.2. (continued on next page)

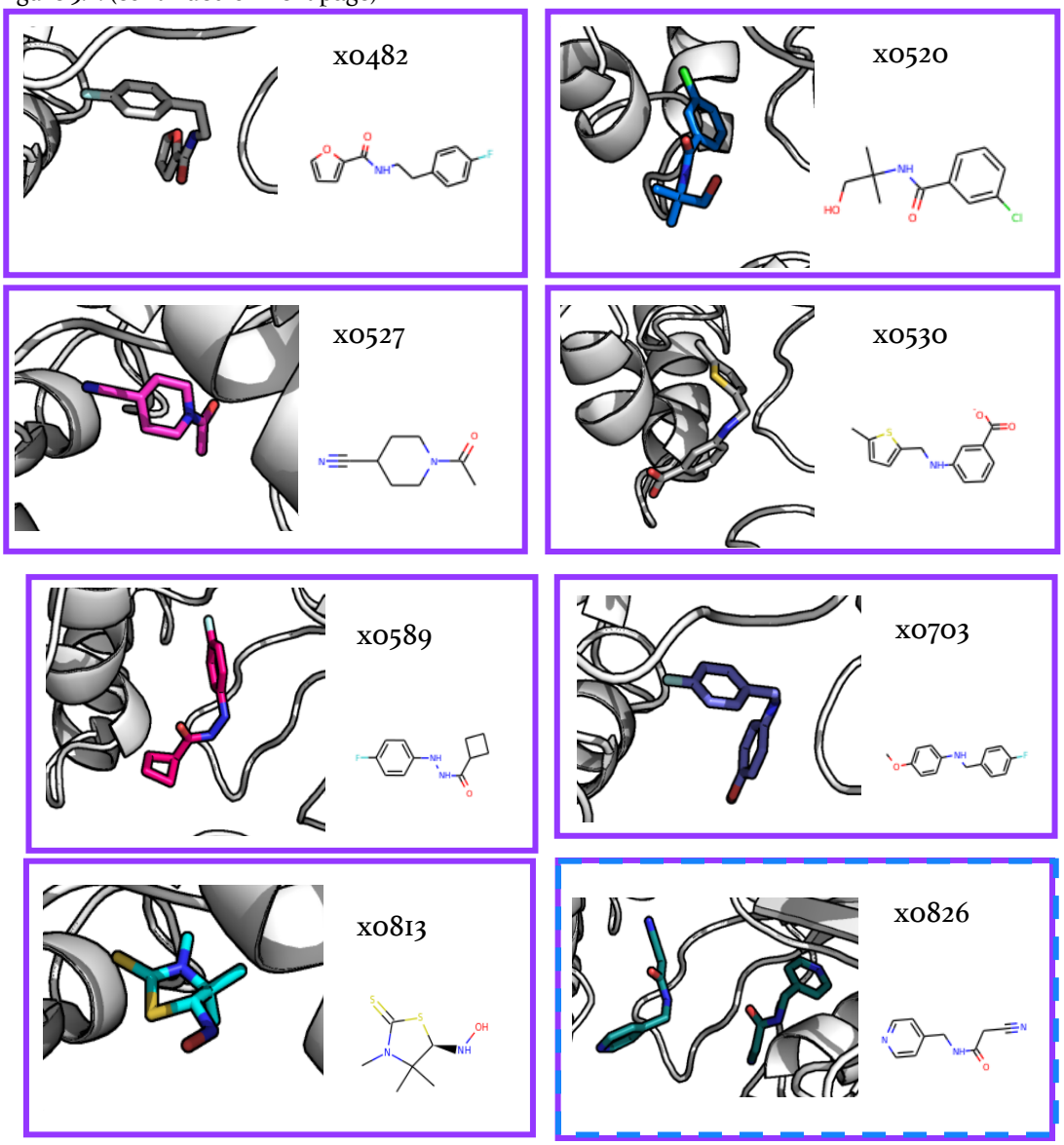



Figure 5.2. *(continued from previous page)* Structures and 2D depictions of each of the high confidence hits for target NUDT22. For each hit the code of the structure refers to an individual crystal structure from the SGC which can be found in Table B.6.1 in Appendix B. The box around each hit indicates which part of the binding site the structure is found in. Purple indicates site 1 of the main binding site, blue indicates site 2 of the main binding site and black indicates a different binding site. These are shown with the protein structures in Figure 5.3. Notably structure x0826 contains two copies of the ligand one bound in site 1 and one bound in site 2 (these sites are labelled in Figure 5.3).

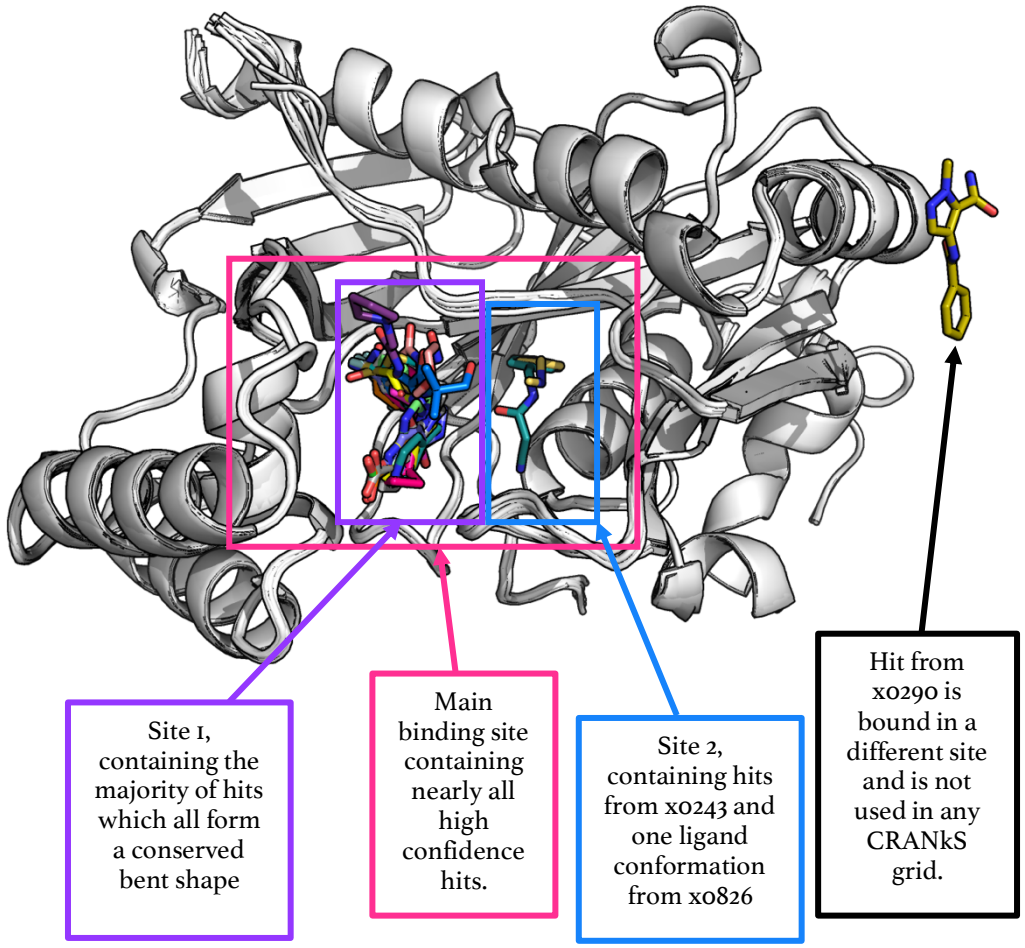



Figure 5.3. Structure of NUDT22 with all high-confidence ligands. The main binding site is shown within a pink box and contains all hits excluding x0290. x0290 is labelled and is bound to a different site. The main binding site can be split in two. Site 1 is shown by a purple box and contains the majority of hits that adopt a conserved shape. Site 2 is shown by a blue box and only contains hits from x0243 and one conformation from x0826.

Candidate compounds to be scored were generated using the XPoise software. This KNIME workflow was developed by Cox *et al.,* (Cox *et al.*, 2016) to use the poised fragments from the library and apply the follow-up reactions *in silico* generating a library of follow-up compounds. Each fragment is deconstructed into two poised synthons. The synthons are then combined computationally with complementary reagents from a chosen catalogue to generate follow-up compounds. The

methodology is shown in Figure 5.4. The catalogue used for this work was the WUXi catalogue - a library of reagents that can be purchased from WUXi (http://www.wuxiapptec.com/chem_catalog_reagents.html).

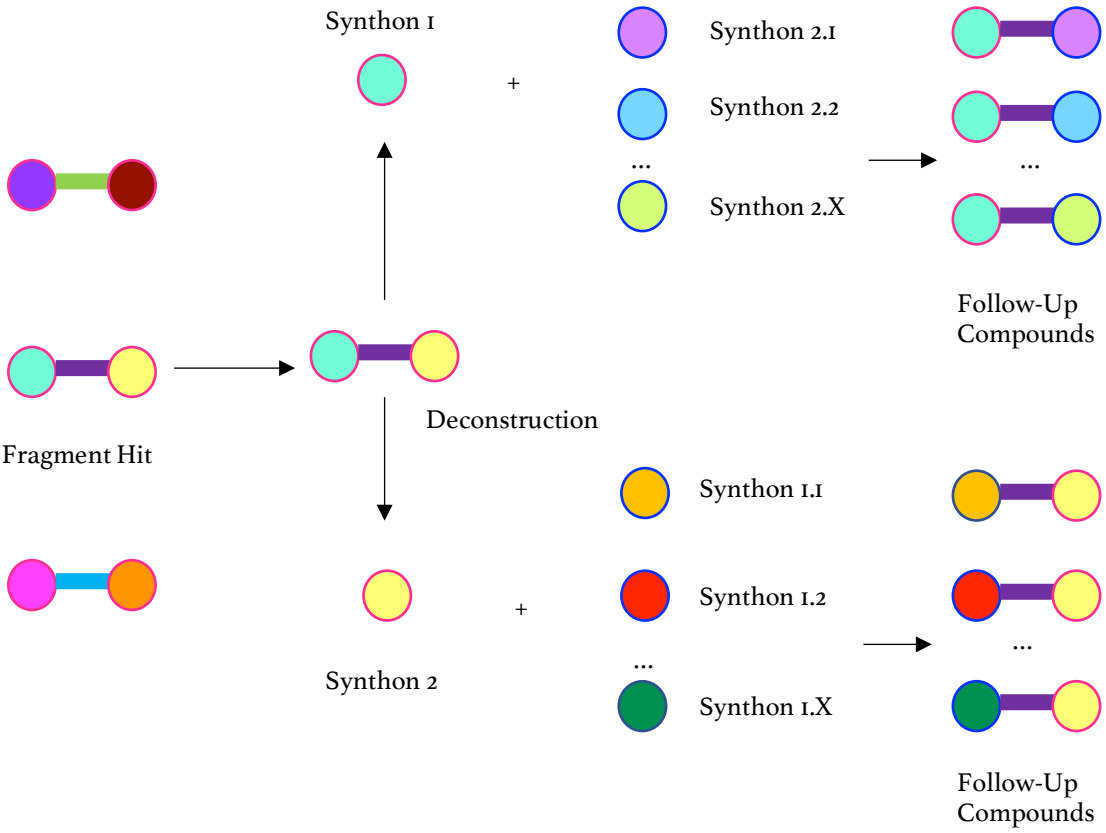



Figure 5.4. XPoise method for generating a library of follow-up compounds. A fragment hit is deconstructed into two synthons. For all the poised reactions the synthon matches, it is computationally reacted with complementary synthons from a reagent library to generate a set of follow-up compounds. Compounds to synthesise using one-step chemistry can then be selected from the follow-up compounds.

For each of the high confidence hits shown in Figure 5.2 both synthons were reacted for follow-up in XPoise using all reactions possible for a synthon. Hits from structures x0290 and x0243 were not included as the ligand in x0290 is bound in an additional binding site, and the ligand from x0243 was only found in site 2. Ligands from structures x0813 and x0329 did not produce any follow-up compounds. A total of 19,028 potential follow-up compounds were generated.

The make-up of those compounds in terms of hits and in terms of reactions used to create the compounds are shown in Figures 5.5 and 5.6. There is a good coverage of the starting fragments in the follow-up library. Both halves of the starting fragment are considered to have originated from that starting fragment, and follow-ups from each half contribute equally to the percentages shown in Figure 5.5. The reaction types used by XPoise to generate follow-up compounds for this library are shown in Table 5.1. The reaction coverage is biased towards reaction 1B (Amide formation with the WUXi reagent as an amine) and 4A (Suzuki coupling with the WUXi reagent as an aromatic halide), due to the make-up of the reagents in the WUXi catalogue. Of the 19,028 molecules, 15,341 compounds were able to be scored by CRANkS. Failures were due to sanitisation problems in RDKit or because no conformers were created that did not clash with the protein.

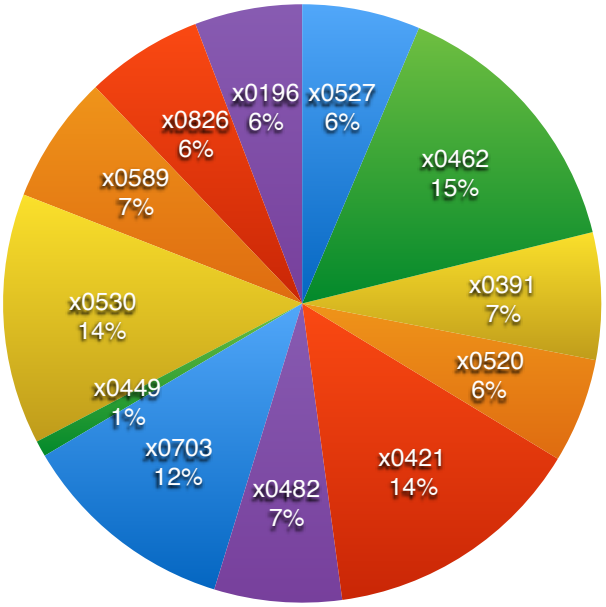



Figure 5.5. Percentage of compounds generated using a particular starting fragment for all the NUDT22A hits for the follow-up compounds calculated by XPoise. There is a relatively even coverage of all starting fragments.

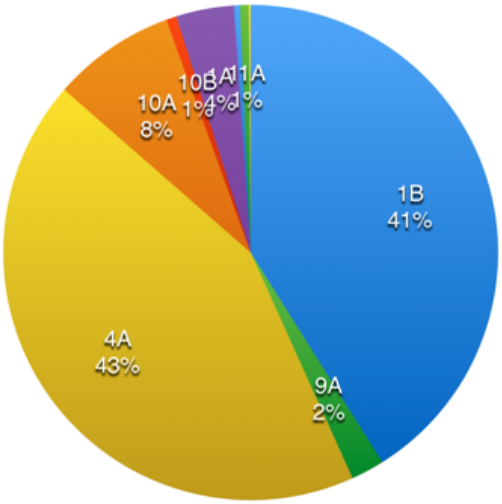



Figure 5.6. Percentage of compounds generated by XPoise formed by each reaction. The reaction space is biased towards reactions: reaction 1B (Amide formation with the WUXi reagent as an amine) and 4A (Suzuki coupling with the WUXi reagent as an aromatic halide).

| Reaction ID | Reaction Name | Varied Synthon |
|---|---|---|
| 1A | Amides | Acids |
| 1B | Amides | Amines |
| 4A | Suzuki Coupling | Aromatic Halides |
| 9A | Ether Coupling | Alcohols |
| 9B | Ether Coupling | Halides |
| 10A | Ester Coupling | Acids |
| 10B | Ester Coupling | Alcohols |
| 11A | Benzimidazole | Acids |
| 11B | Benzimidazole | Diamines |

Table 5.1. Reactions used by XPoise to create a follow-up library using the high confidence hits for target NUDT22 and the WUXi catalogue.

### *5.2.3 Selection Set One – PLIF Diversity*

The first selection only involved hits in site 1 of the main binding site in Figure 5.3. This is the site where the majority of hits are bound with a conserved shape and thus was the focus for follow-up. Therefore, all hits excluding x0290, x0826 and one conformation of the ligand hit from x0243 found in site 2, were used to construct CRANkS grids with the algorithm described in Chapter 3.

The first selection set was chosen based on two hypothesises. The first is that the Interaction Score could be used to show enrichment: compounds that form interactions that the hits are also calculated to form, are more likely to bind. These compounds will have low Interaction Scores as the interaction patterns will overlap with the interaction grids. The second is that the selected compounds should cover a wide range of the interaction space, i.e. form different interactions to each other. This is to explore the space as widely as possible so that the campaign is not restricted to particular follow-up series at this early stage of searching for a suitable chemical probe. I hypothesise that by maximising the diversity of the protein-ligand fingerprints (PLIFs), a wide coverage of the possible interaction space will be achieved.

The CRANkS grids were thus constructed and each follow-up compound from XPoise was calculated to have an Interaction Score, Element Score, and Pharmacophore Score. To select compounds with a low Interaction Score the compounds were ranked by ascending Interaction Score, and the 500 compounds with the lowest Interaction Scores were taken forward. For each of these 500 follow-

up compounds the CRANkS algorithm generated 100 conformers, which could form different interaction patterns. Thus, PLIFs were generated for each conformer for each of the 500 follow-up compounds. The PLIFs were generated by ordering all interactions formed by any conformer. A bit of the fingerprint is assigned to each interaction. The fingerprint is then generated by assigning 1 to the bit if the conformer forms the interaction, and 0 if the conformer does not.

To select a subset of follow-up compounds for synthesis, conformers were selected to maximise the PLIF diversity using the *MaxMinPicker* algorithm in RDKit (Ashton *et al.,* 2002). This is a computationally efficient way of selecting a smaller subset from a larger set of molecules to maximise the diversity. The algorithm requires the number of molecules to be selected and a distance matrix specifying the similarity between the molecules. In this case the distance matrix was generated using the Tanimoto similarity (described in Section 1.3.1) calculated using the PLIFs for each of the conformers.

Different numbers of conformers were selected from the 500 compound set in the way described. Table 5.2. shows the number of conformers selected and the corresponding number of compounds selected. Figure 5.7 shows the relationship between the number of conformers selected to maximise PLIF diversity and the subsequent number of compounds selected. An interesting behaviour can be observed. The number of compounds does not increase linearly with the number of conformers. This indicates that some molecules have conformers generated that cover more interaction space than others. This is likely to be inherently linked to the conformer generation step. Compounds with a large overlap with the ligands are

placed with a large maximum common structure. For these compounds there are therefore fewer bonds and atoms for which a new conformation is generated causing few conformers to be generated that have a large difference in orientation. This highlights a potential problem with the conformer generation step of the CRANkS algorithm – there is no measure of the difference between the conformers generated, currently all 100 could be identical.

| Number of Conformers Selected | Number of Compounds Covered by Selection |
|---|---|
| 50 | 30 |
| 100 | 41 |
| 150 | 55 |
| 200 | 66 |
| 300 | 104 |
| 500 | 119 |
| 1000 | 121 |

Table 5.2. The number of conformers selected to maximise PLIF diversity from a 500 compound set with the lowest Interaction Scores from the XPoise follow-up libraries. The resulting number of compounds selected are also shown. There is a non-linear relationship between the number of conformers selected and the subsequent number of compounds, indicating some compounds have conformers generated that cover a large amount of interaction space and others do not. This is shown in Figure 5.6.

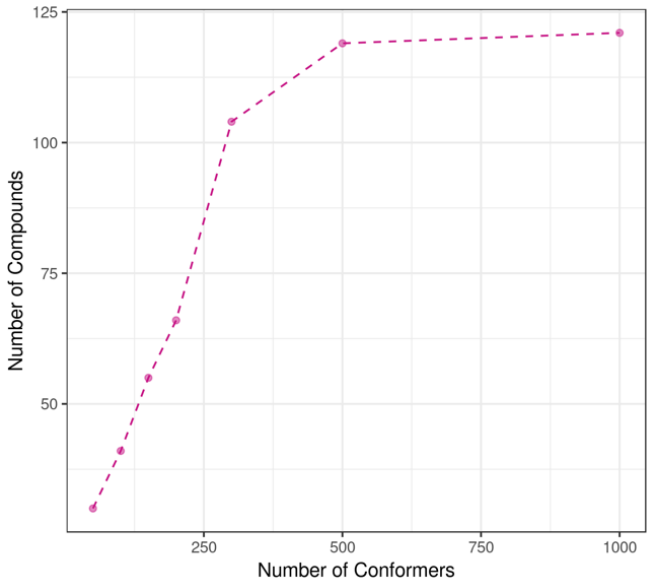



Figure 5.7 The number of conformers selected to maximise PLIF diversity against the consequent number of compounds selected. There is a non-linear relationship.

An example of conformers generated for two compounds by this method are shown in Figures 5.8 and 5.9. Compound 20313 was selected as part the subset generated to maximise the PLIF diversity of 50 conformers, and forms part of the 30 compound set shown in Table 6.2. For compound 20313 six separate conformations are covered by the 100 conformers generated by the CRANkS algorithms. Conformations were considered to be different if the RMSD between the conformations was greater than 0.005 Å. The separate conformations are shown in Figure 5.8 A and the proportions of the 100 conformers generated adopting each conformation is shown in Figure 5.8 B. For compound 2909, which was not selected by this method for the 30 compound set, only two distinct conformations were formed and the number of the 100 conformers generated forming each conformation are shown in Figure 5.9 A and B respectively.

Compound 20313 was selected when using this method, for the subset selected to maximise the diversity of 50 conformers. The conformations formed are shown in Figure 5.8 A. This is typical of compounds selected by this method – there are three rotatable bonds for which conformers can be generated. The remaining bonds are held in place by the constrained conformer generation due to overlap by one of the ligands. The 100 conformers generated by CRANkS only generate six separate conformations. The proportions of the conformers adopting each of the orientations is shown in Figure 5.8 B. These proportions will have an effect on the scores calculated by the CRANkS algorithm as the total score for the compound is the average of the score for each conformer over the 100 conformers generated. Additionally, a mixture of stereoisomers is calculated which is discussed in further detail below. This indicates an issue with the current method as these meaningless

proportions could be changing the scores of compounds. The method should be adapted so that one conformer for each conformation is taken forward rather than the full 100 conformers.

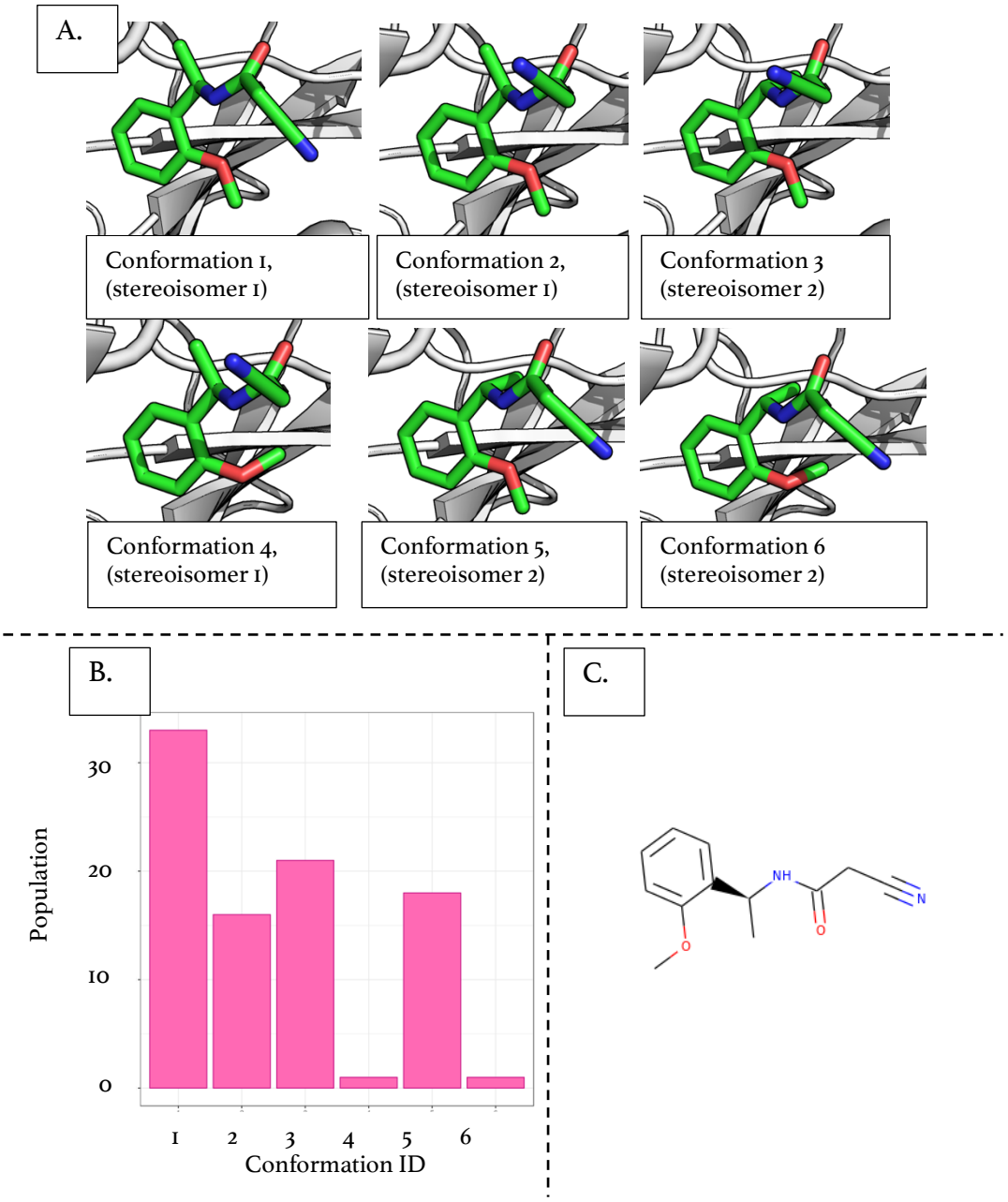



Figure 5.8 Conformations adopted by the 100 conformers generated for compound 20313. A. shows each of the separate conformations generated. B. shows the proportion of the 100 conformers adopting each of the conformations and C. shows the 2D structure of the compound. A mixture of stereoisomers is calculated with the stereoisomer labelled in terms of the chiral centre highlighted by * in C. The proportion of the conformations adopted by the 100 conformers will have an effect on each of the CRANkS scores, as the scores are an average over the 100 conformers. This highlights an issue with the conformer generation step.

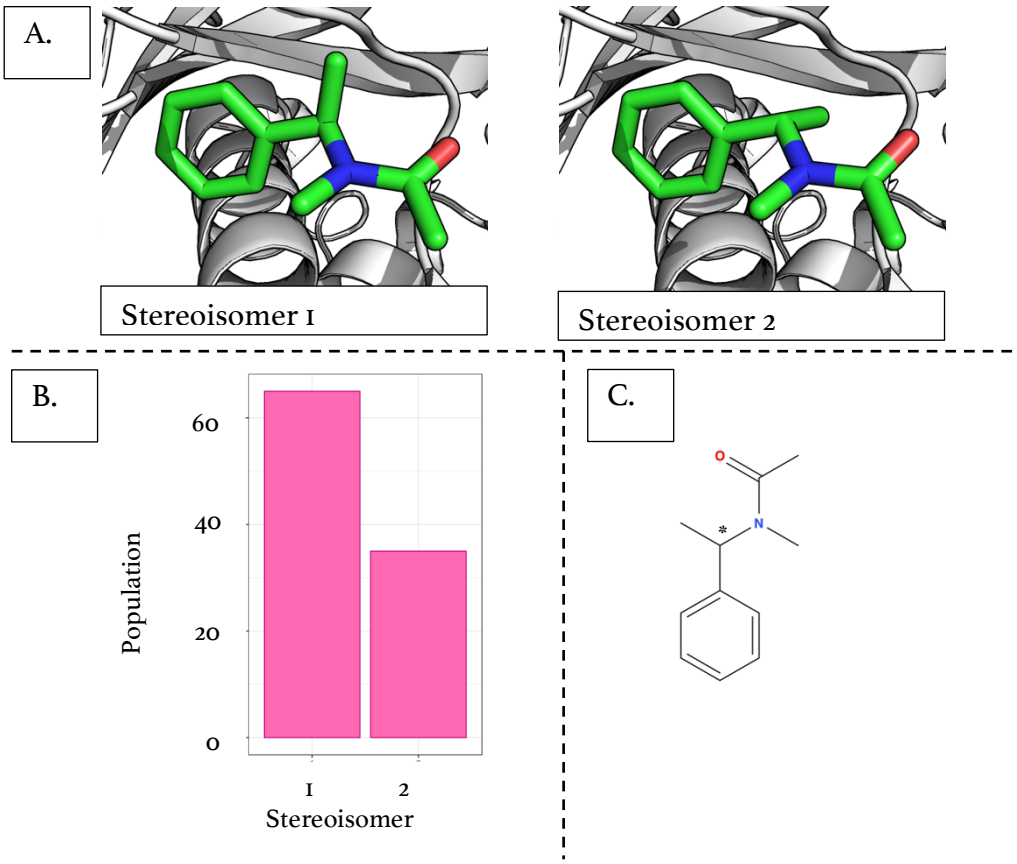



Figure 5.9 Conformations adopted by the 100 conformers generated for compound 2909. A shows each of the separate conformations generated – these are the two stereoisomers for the chiral centre highlighted in C. B. shows the proportion of the 100 conformers adopting each of the stereoisomers and C. shows the 2D structure of the compound. There is only one rotatable bond that is not held in place by the constrained conformer generation. Despite this, 100 conformations are generated by the algorithm which is computationally inefficient.

Figure 5.9 shows that for compound 2909 there is a high number of atoms in the maximum common structure between the compound and one of the ligands. This only allows the orientation of one methyl group to change conformation as all other groups are head in place by the nature of the constrained conformer generated. Two stereoisomers are generated with one conformation of each. This highlights an issue with the algorithm – a racemic mixture of stereoisomers is calculated for a candidate compound if the stereochemistry is not specified. This also shows the computational inefficiency of the current computational method. 100 conformers are generated

despite there only being a single rotatable bond. There is a need for the conformer generation to cluster generated conformers to ensure that the conformer space is sampled sufficiently and efficiently.

For the set of compounds chosen when 100 conformers are selected using this method, the coverage of starting hit is shown in Figure 5.10. Unlike for the whole follow-up library the selected compounds are biased towards compounds originating from x0482 and x0520. The diversity of starting fragments is not maintained. Coverage of the starting fragment space is important at this stage of the process of finding a chemical probe. Limiting the chemical space of follow-ups at this early stage could limit the possibility of finding the most amenable candidate. Due to the inherent bias of the selection to compounds with a low maximum common substructure with the hits and lack of coverage of starting fragment this selection method was not taken forward.

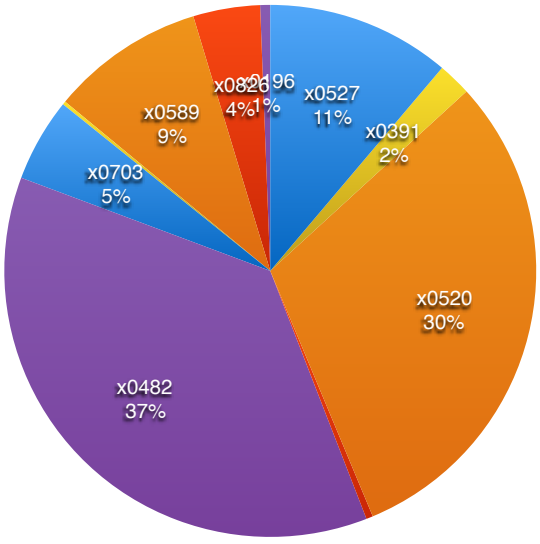



Figure 5.10. Proportions of compounds generated from each starting fragment using selection set 1. Compared to the proportions of starting fragments from the whole follow-up library (Figure 5.4) there is now a large bias towards starting fragments x0482 and x0520 and there is not a representative coverage of the starting fragment space.

### *6.2.4 Selection Set Two – Starting Fragment Diversity*

The first selection set described in Section 6.2.3 was found to lack coverage of the starting fragments. Using follow-up compounds originating from each starting fragment is important at an early stage of drug discovery or chemical probe development, in order to cover as much chemical space as possible. To ensure the selection of compounds would maintain starting fragment diversity (rather than selection set 1 which showed starting fragment bias) the same ratio of starting fragments as the full follow-up library is imposed on the subset for selection subset 2. The proposed selection method is shown in Figure 5.11.

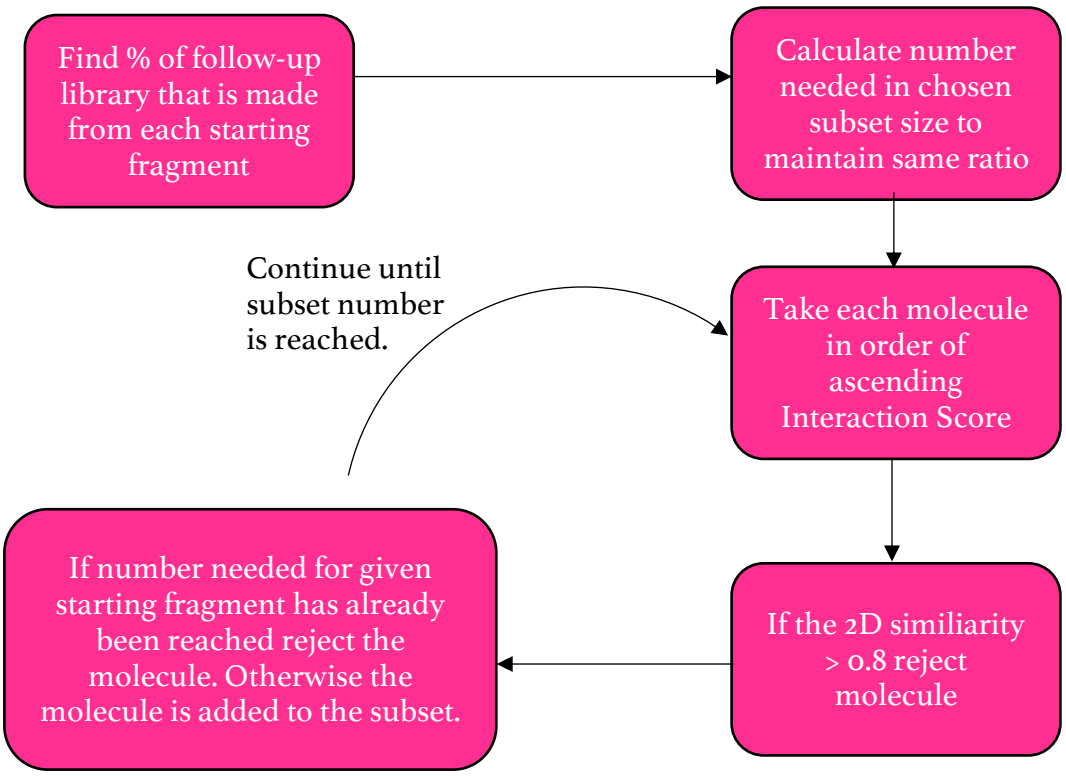



Figure 5.11. Selection method two for selecting follow-up compounds for NUDT22. The ratio of starting fragments in the follow-up library is imposed on the subset to ensure starting fragment diversity. Molecules are selected to prioritise a low Interaction Score, thus selecting compounds that form interactions already observed in the fragment hits. A 2D similarity cut-off is applied to maintain molecular diversity.

The number of compounds to be made *i.e.* the size of the subset, is decided first. The number of compounds needed originating from each fragment hit is calculated to maintain the same ratio of the starting fragments as the follow-up library. For example, for the fragment hit from structure x0530, 14% of the follow-up compounds in the XPoise generated library were generated from this starting fragment. As a result, for a subset size of 50 compounds selected by this method, seven of the compounds should have been generated by XPoise from this starting fragment.

The molecules are then ordered by ascending Interaction Score. I hypothesised that the lower the Interaction Score, the more interactions are formed that match interactions formed by the hits, and the more likely the compound will bind to the target. Each molecule is then taken in turn. Given the starting fragment for that compound, if the number of required molecules originating from that starting fragment has not been reached and the average 2D similarity between the compound and the molecule is less than 0.8 the compound is added to the subset. This continues until the number of molecules required has been selected. The average 2D similarity is the average Tanimoto similarity (Section 1.3.1) between the Morgan fingerprint (using a radius of 2, see Section 3.2.5.1) for the compound in question, and the Morgan fingerprints of the compounds already chosen to be in the subset. This should ensure the molecular diversity of the compounds in the set. A 2D similarity cut-off of 0.8 was chosen as this is conventionally chosen as a cut-off for similar compounds due to work on QSAR and related similarities between compounds and activity (Jasial *et al.,* 2016).

Three subsets of 50, 100 and 200 compounds were respectively selected for analysis. First the pairwise 2D similarity of each compound to all other compounds in the set is plotted as a histogram and shown in Figure 5.12. This is indicative of the molecular diversity of the subset. A random selection of compounds of the same number as the selected subset is repeated 100 times and the average pairwise 2D similarity across all the compounds averaged over the 100 subsets is plotted as a blue line. This allows comparison of the molecular diversity of the selected subset using the method described here, with a subset of the same number of compounds selected at random. The histograms show that the majority of compounds are less similar to each other in the subsets than if selected randomly from the follow-up library. The selected subsets by this method are therefore covering a wider range of molecular diversity than if the compounds were chosen at random. Additionally, the histograms have approximately the same range, indicating that the molecular diversity is maintained across subset sizes.

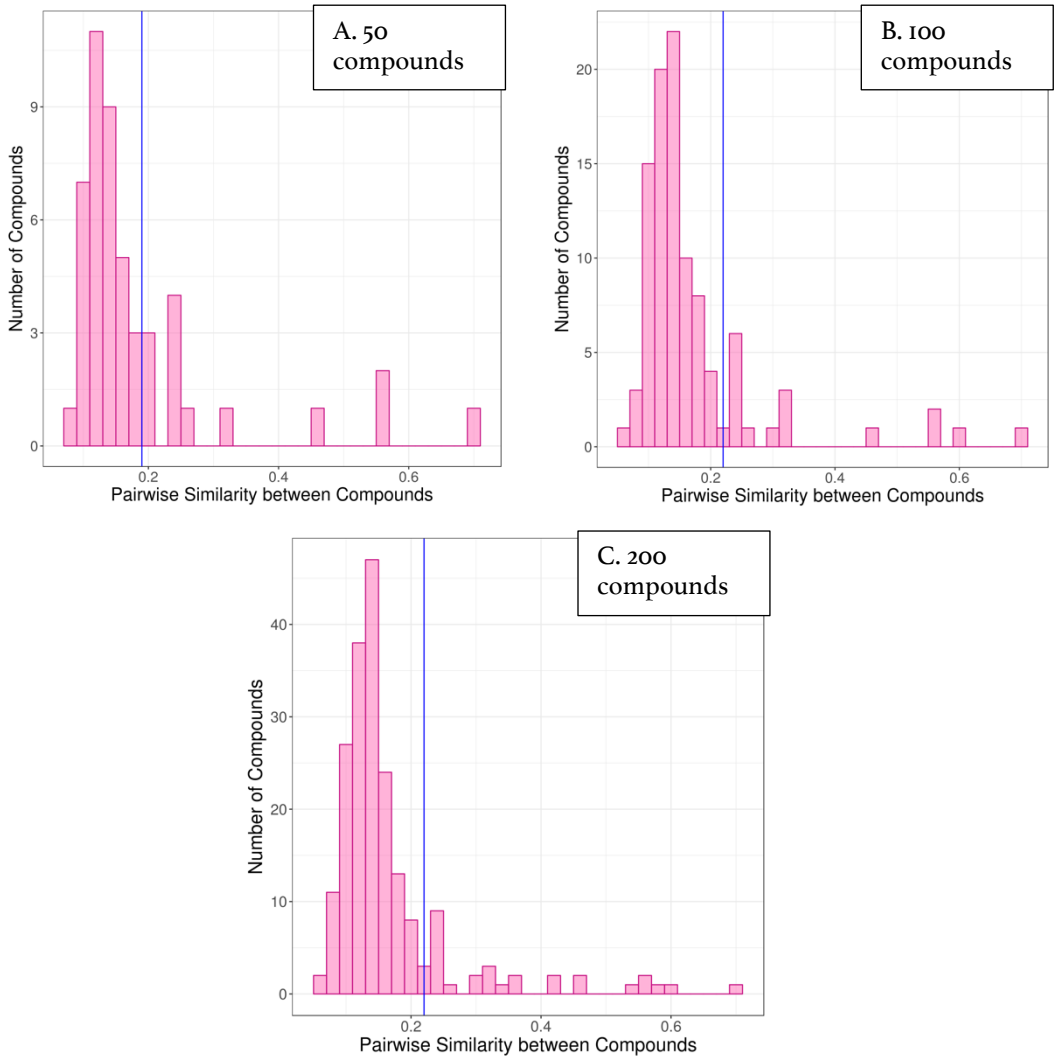



Figure 5.12. Distributions of the pairwise 2D similarities between each of the compounds in the subset and all other compounds in the subset. This is shown for subsets of 50 (A), 100 (B) and 200 (C) compounds. The blue line is the average over 100 randomly selected subsets from the follow-up library of the same size of the pairwise 2D similarity between each compound in the subset and all other compounds in the subset. The compounds in the subset are less similar to each other than on average for randomly selected compounds.

Although the starting fragment diversity has been maintained this does not ensure synthetic route diversity. Figure 5.13 shows the distribution of reactions for each of the subsets, and the corresponding reaction for each reaction id is shown in Table 6.1. The reaction diversity is approximately maintained across the subset sizes. There is a slight reduction in the proportions of the less dominant reactions, for example 10 A (ester coupling), but this is not significant.

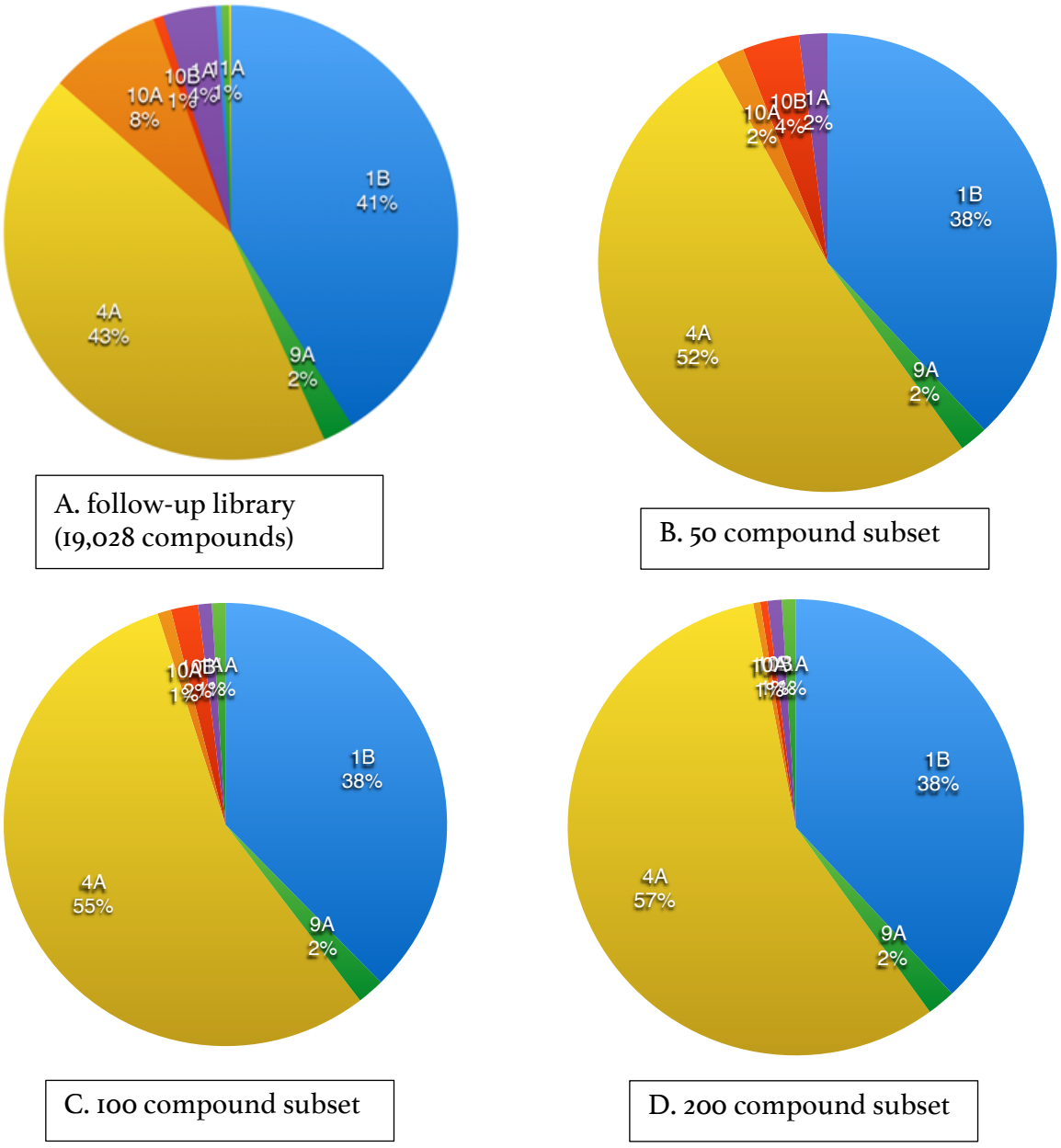



Figure 5.13. Proportion of each reaction used to create compounds in each of the subsets. A shows the proportions for the whole follow-up library, B for the 50 compound subset, C for the 100 compound subset and D for the 200 compound subset. The coverage of reaction space is approximately maintained across the subsets.

Consequently, I proposed that the 100-compound subset be synthesised, and the required reagents purchased. The whole set of compounds are shown in Figure A.5.1 in Appendix A. The compounds were studied by a synthetic chemist to ensure the

synthetic tractability. 11 compounds were removed from the subset. These are shown in Figure 5.14. The compounds coloured in red are electron rich anilines and consequently the reactivity of these compounds may cause problems. The compounds coloured in yellow were removed due to the reactive heterocycles with halogen atoms. The compounds coloured in blue are thioethers and consequently were removed. The compounds coloured in green were removed because the compounds were determined to be too hindered to be able to be synthesised. The compound coloured in pink was removed because it was too small to be of interest and to yield further information about the binding site.

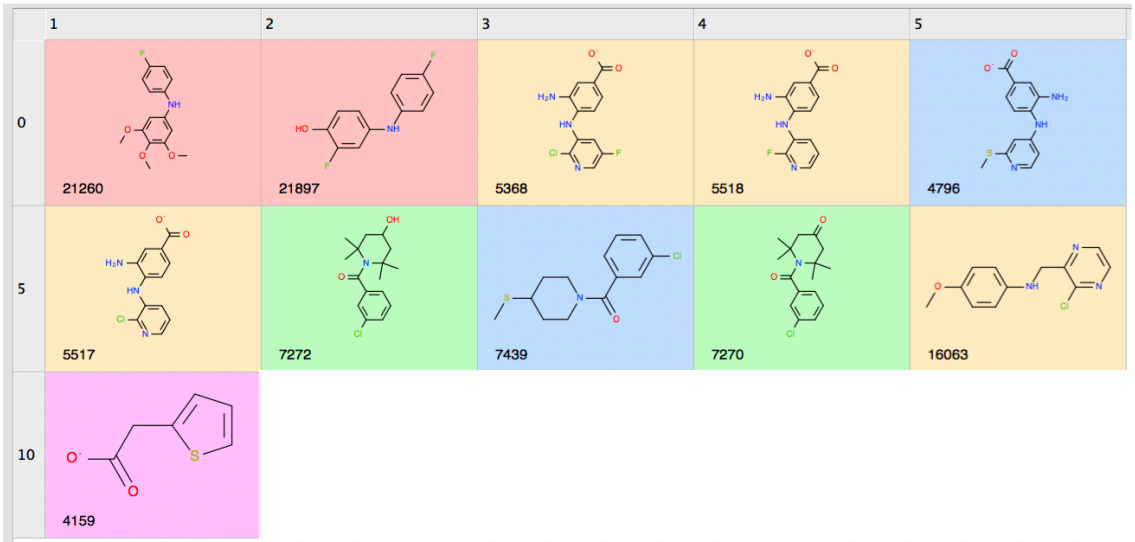



Figure 5.14. Compounds rejected from synthesis by a synthetic chemist due to reactivity or inability to synthesise. Compounds coloured in red are electron rich anilines, in yellow are reactive heterocycles and in blue are thioethers, all of which were removed due to reactivity. Compounds coloured in green were considered to be too hindered to synthesise. The compound coloured in pink was removed due to its small size.

However, now PLIF diversity has been completely neglected in this selection of compounds. The diversity of the protein-ligand interactions formed by each of the compounds chosen in the subset, for a subset size of 25, is shown in Figure 5.15. The *x*-axis represents each of the interactions of the interaction grid and the *y*-axis represents each of the compounds. A subset of 25 is shown as the number of compounds able to be purchased was reduced. For each compound the square is coloured if one of the conformers forms that interaction. The intensity of the colour refers to the number of conformers forming that interaction. It is clear that the molecules form similar interactions and there is not a wide coverage of the interaction space. On inspection of the conformations in the binding pocket, this was discovered to be because nearly all the compounds were being placed with a maximum common substructure from the same fragment, rather than the individual starting fragments. This causes the chosen compounds to have similar interaction patterns.

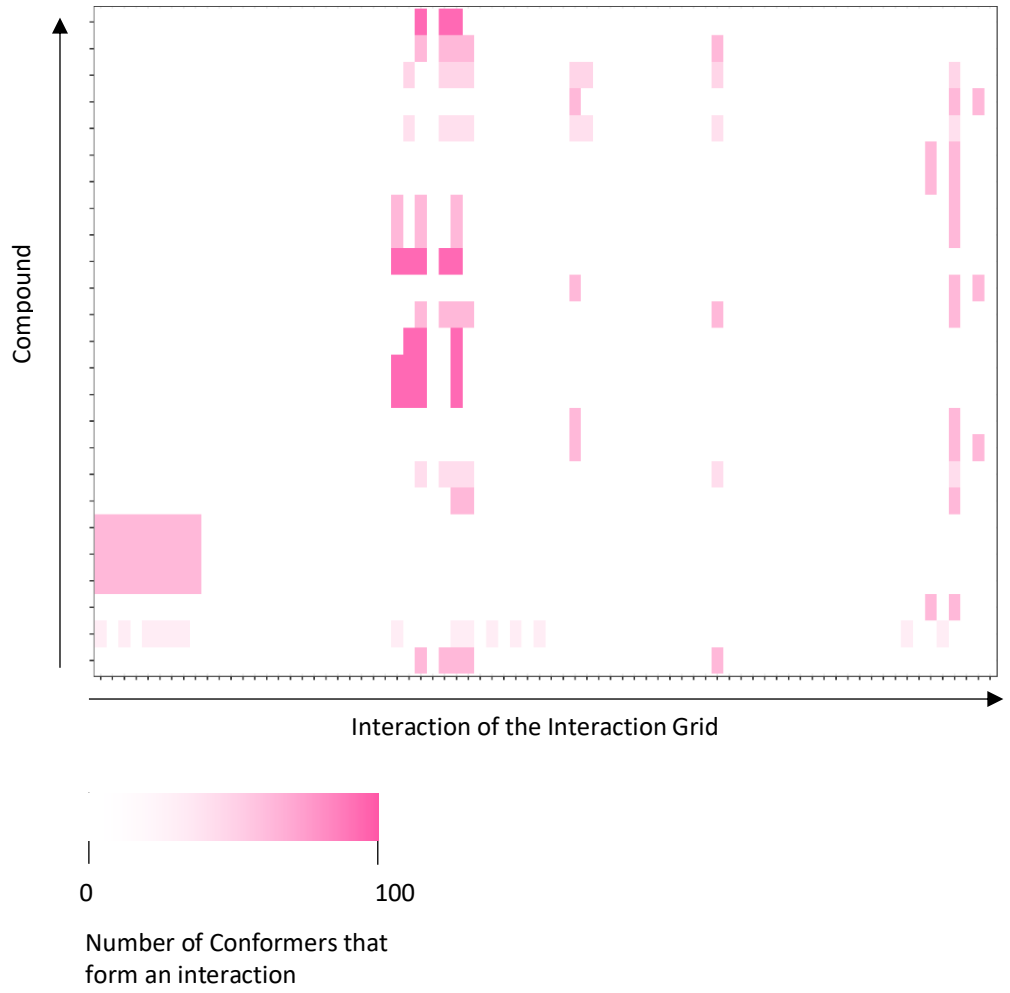



Figure 5.15 Plot showing the interactions formed by each of the compounds in a 25 compound set. Each interaction in the interaction grid is shown on the *x*-axis and each compound is on the *y*-axis. The grid is coloured if the compound forms that particular interaction. The intensity of the colour reflects the number of conformers forming this interaction. Although the ratio of starting fragment is conserved to match the ratio of starting fragment in the follow-up library there is a lack of diversity in terms of interactions formed.

The CRANkS algorithm does not ensure that a given candidate compound is placed in the binding site by its starting fragment. A candidate compound is placed by the fragment with which it has the largest maximum common substructure – its placement fragment. Consequently, rather than select the compounds to keep the ratio of the starting fragment equal to that of the follow-up library, the subset was

reselected to keep the proportions of the placement fragments of the subset selected equal to the ratio of starting fragment of the follow-up library All other parts of the procedure remain the same. The resulting interaction patterns of the 25 subset is shown in Figure 5.16. This selection gives a larger coverage of interaction space with more of the interactions of the grid covered by the selected compounds. The number of interactions formed found in the grid increases from 20 (27% of the interactions in the grid) to 36 (48% of the interactions in the grid). A 25 molecule subset was reselected using this method and was taken forward (shown in Appendix A, Figure A.5.2).


Figure 5.16. Plot showing the interactions formed by each of the compounds in a 25 compound set. This set was constructed by imposing the ratio of starting fragments found in the follow-up library, on the ratio of fragments placing the compounds in the subset. Each interaction in the grid is shown on the *x*-axis and each compound is on the *y*-axis. The grid is coloured if the compound forms that particular interaction. The intensity of the colour reflects the number of conformers forming this interaction. There is a more diverse coverage of interactions than in Figure 5.15.

### *6.2.6 Selection Subset 3 - Novel Interactions*

Subsequent to the presentation of selection subset 2 to the follow-up team, issues were raised about the lack of novel interactions that could be selected by the method. By minimising the Interaction Score to select compounds that form the same interactions formed by the fragment hits, it is likely compounds are being prioritised that do not form any novel interactions. Each of the compounds in selection subset two were inspected within the binding site. An example of a typical predicted interaction pattern for one of the compounds is shown in Figure 5.17. All conformers for the compound are shown. The compound forms two interactions that are not formed by the placement fragment (the maximum common substructure between the compound and any of the hits) – hydrophobic interactions that are labelled by arrows in Figure 5.17 B. This is typical of compounds in the set with many only forming hydrophobic interactions that were not formed by the placement fragment. A number of compounds formed no new interactions.

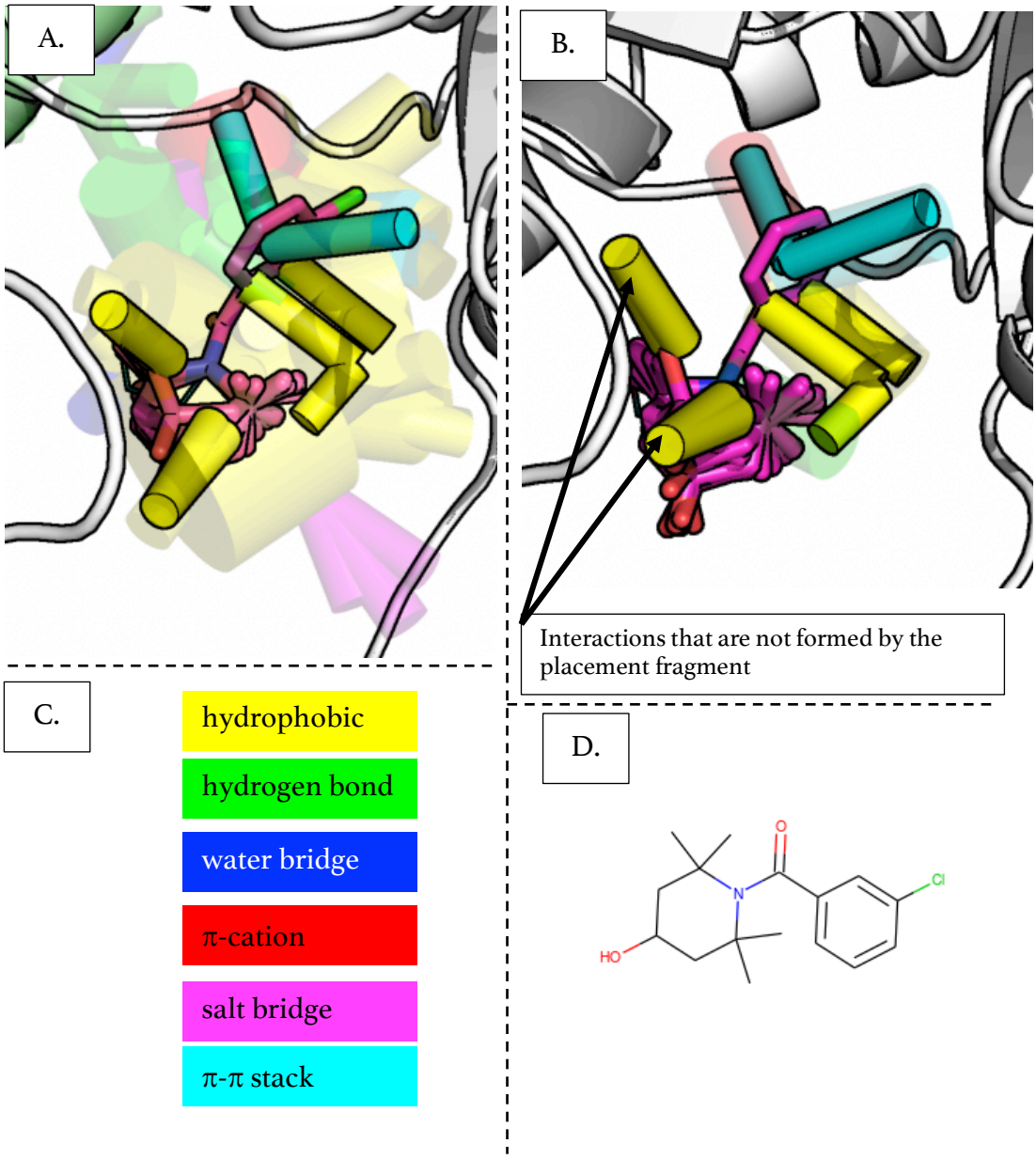



Figure 5.17. Calculated interactions for conformers for a compound for the 25-molecule subset of subset selection 2. This is typical of compounds in the set. A shows the interactions formed by the conformers in opaque cylinders and the whole interaction grid in transparent cylinders The width of the cylinders in the interaction grid indicate the count of the interaction grid. The interactions formed by the compound all overlap with interactions in the interaction grid - no interaction are novel compared to the grid. B shows the interactions formed by the conformers in opaque cylinders and the interactions formed by the placement fragment as transparent cylinders. The compound fulfils the interactions of the placement fragment and only forms hydrophobic interactions that are novel compare to the parent compound which are labelled in B. C shows the key for the types of interactions in A and B. D shows the 2D structure of the compound.

This highlights that the Interaction Score could be used in two different ways. The first explored in the previous selection methods hypothesises that molecules that are calculated to form interactions that are similar to the hits are more likely to be active. Therefore, I hypothesised that minimising the Interaction Score and using this to select compound will prioritise active compounds. The second is the molecules that are calculated to have a high novelty score with form novel interactions and allow further exploration of interaction space.

To prioritise compounds that form novel interactions the Interaction Score was partitioned so that the interactions formed by the placement fragment (the maximum common substructure between a candidate compound and any fragment hit) are not included in the score. This score is referred to as Interaction Score (No Frag). Molecules with a high Interaction Score (No Frag) will form interactions that are different to those formed by the fragment hits and so are exploring new interaction space. By maximising the Interaction Score (No Frag) compounds should be selected that form novel interactions with the target.

I therefore proposed to select two sets of compounds for synthesis based on these two hypotheses. The first set is chosen to minimise the Interaction Score to prioritise compounds that form similar interactions to the fragment hits and therefore are thought to be more likely to be active (I will refer to this as subset 3.1). The second set is chosen to maximise the Interaction Score (No Frag) to select compounds that form novel interactions compared to the fragment hits (I will refer to this as subset 3.2). This should yield compounds that bind and novel compounds in terms of interactions.

The entire follow-up library was therefore rescored to calculate the Interaction Score (No Frag). Two subsets were selected using the SQUONK diversity picker to either minimise the Interaction Score or maximise the Interaction Score (No Frag). The SQUONK diversity picker was used to ensure 2D molecular diversity of the subset. SQUONK is a workflow management system being trialled by the XChem group at Diamond Light Source (*http://github.com/InformaticsMatters/squonk*). It works in a similar way to KNIME with nodes that link together to execute algorithms. However, it more easily allows users to create their own nodes using DOCKER packages (Merkel, 2014). The diversity picker should select compounds from a library to create a subset of a certain size, either maximising or minimising a given variable (in this case either the Interaction Score or Interaction Score (No Frag)) whilst maintaining 2D molecular diversity.

Two subsets of 25 compounds was selected by this method. To ensure the coverage of interactions in the interaction grid, as these interactions were not considered in the selection of these compounds, Figures 5.18 and 5.19 show which interactions of the interaction grid are formed by each compound. Figure 5.18 shows the interaction diversity for the subset chosen to minimise the Interaction Score, i.e. subset 3.1. Of the 75 interactions of the interaction grid, 34 are covered by the subset. This is a small reduction of 2 interactions (2.7%) from subset 2. Figure 5.19 shows the interaction diversity for the subset chosen to maximise the Interaction Score (No Frag) i.e. subset 3.2. A similarity coverage of interaction space is achieved with 36 interactions from the interaction grid covered by the 25 compound subset. This is despite no

imposition of proportions of placement fragment on the selection. The sufficient coverage is likely to be achieved by the 2D molecular diversity of the subset.

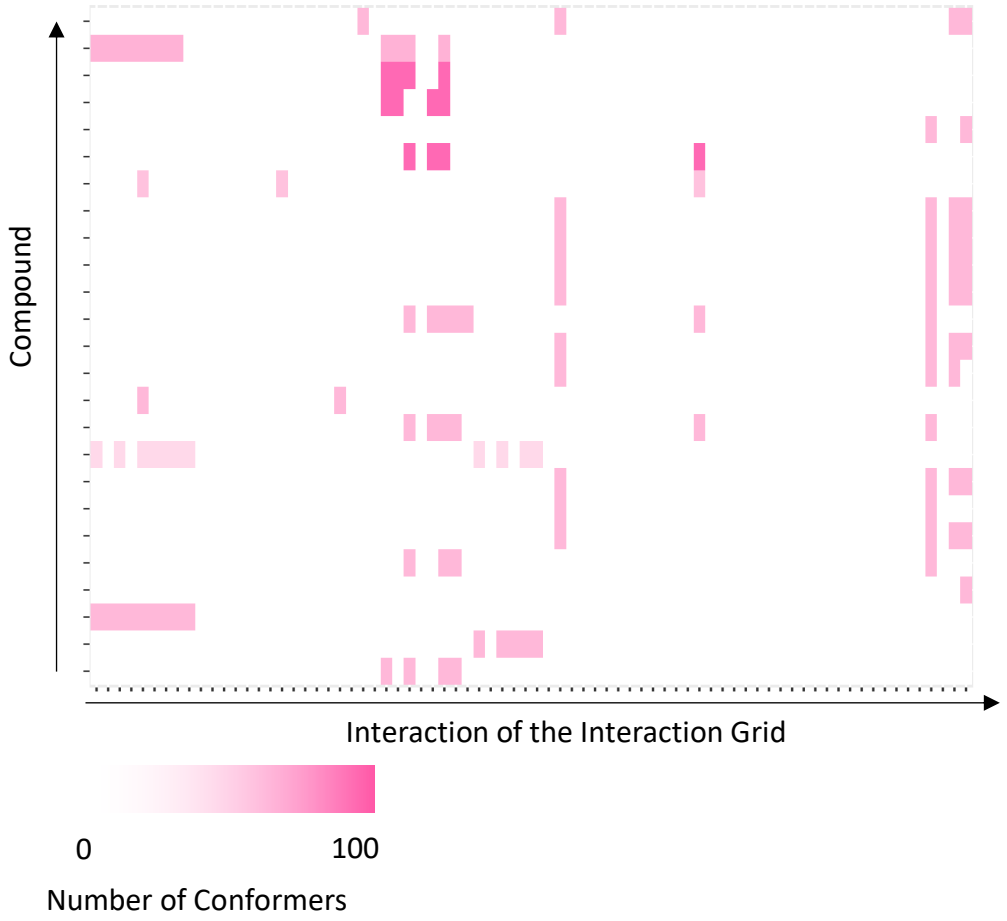



Figure 5.18. Plot showing the interactions formed by each of the compounds compound set 3.1. This was selected using the SQUONK Diversity Picker to minimise the Interaction Score. The coverage of interactions is similar to the coverage of subset 2 (34 interactions for subset 3.1 and 36 for subset 2).

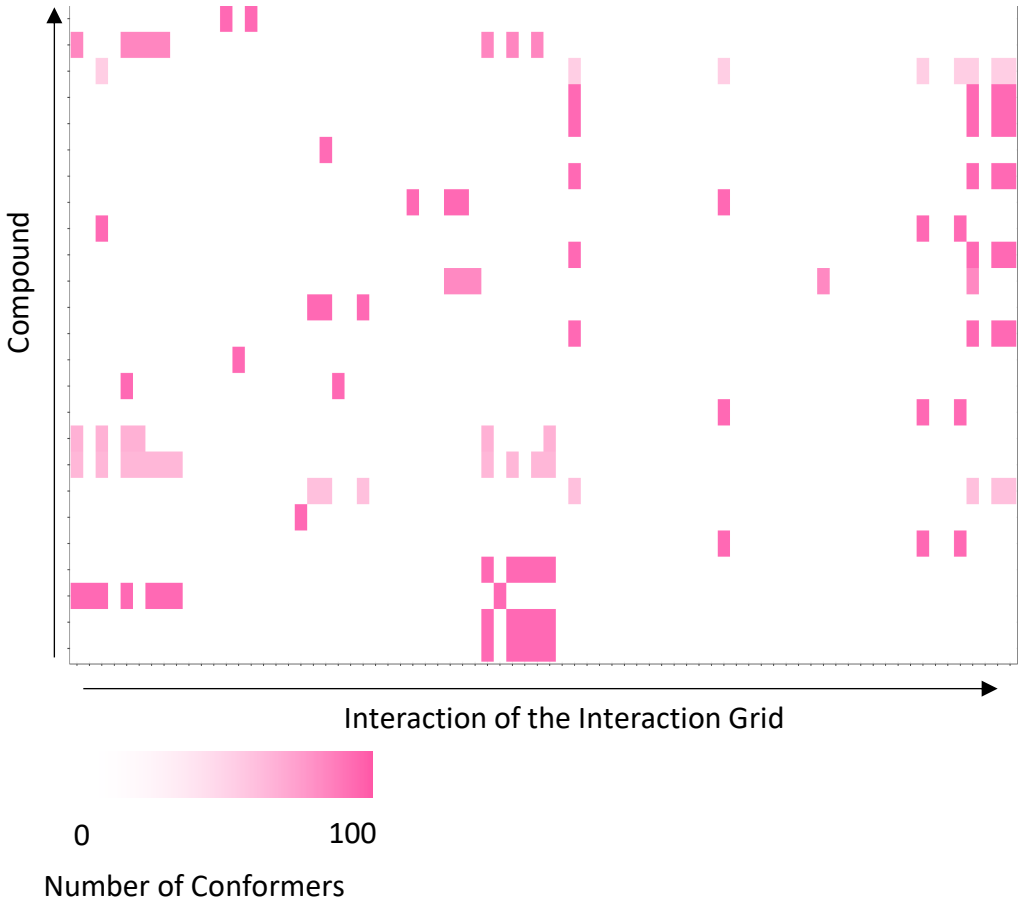



Figure 5.19. Plot showing the interactions formed by each of the compounds compound set 3.2. This was selected using the SQUONK Diversity Picker to maximise the Interaction Score (No Frag). The coverage of interactions is similar to the coverage of subset 2 (37 interactions for subset 3.2 and 36 for subset 2).

To ensure novel interactions are selected by subset 3.2, the number of novel interactions formed by each of the compounds in the subset was calculated. The interactions are novel if the interactions are not found in the interaction grid i.e. if the interactions were not formed by any of the fragment hits. All compounds are found to form novel interactions with some compounds even forming over nine interactions. Consequently, both subsets of 25 compounds selected by these methods (subsets 3.1 and 3.2) were also proposed for synthesis (Appendix A, Figures A.5.3 and A.5.4).

### *5.2.6 NUDT22 Conclusions*

A number of methods for the selection of follow-up compounds for target NUDT22 are explored in this section. The compounds were selected from a follow-up library using the XPoise KNIME workflow which generates follow-up compounds *in silico* using commonly-used reactions from poised fragment hits (Cox *et al.,* 2016). Table 5.3 summarises the selection methods described and what each method looks to achieve.

| Characteristic | Measure | Subset 1 | Subset 2 | Subset 3.1 | Subset 3.2 |
|---|---|---|---|---|---|
| Compounds that are active | Minimised Interaction Score | | | | |
| Compounds that cover the interaction space of the hits | PLIF diversity or placement fragment diversity | | | | |
| Compounds that form novel interactions | Maximised Interaction Score | | | | |
| Compounds that cover a wide-range of chemical space of the follow-up library | 2D similarity using Morgan Fingerprints | | | | |

Table 6.3. The characteristics required for the compounds in the subset are shown in the first column and the measure of those characteristics is shown in the second. The characteristics covered by each subset is shown by a green cell.

Compounds that are active are selected by minimising the original Interaction Score. I hypothesise that compounds calculated to form the same interactions as the

fragment hits are more likely to be active. Subsets 1, 2 and 3 looked to achieve this by three different methods.

The first method preselects a set of 500 compounds with the lowest Interaction Scores. Molecules are then further sub-selected to maximise the PLIF diversity of the conformers. This method was rejected due to the lack of coverage of the starting fragments used to generate the subset of compounds. This method also highlighted an issue with the conformer generation step of the CRANkS algorithm. Some compounds were calculated to have conformers that covered a large amount of interaction space and others did not. Currently there is no measure on how different two conformers generated are to each other, and no conformers are rejected based on similarity to the other conformers generated for a particular compound. This could bias results as some compounds could have 100 identical conformers, whereas others could have 100 conformers that are very different to each other in terms of the RMSD between the conformers.

Subset 2 also selected compounds with a high coverage of the interaction space of the hits (Figure 5.16). However, the subset maintained molecular diversity by ensuring the ratio of fragment hits used to place the compounds within the binding site (the maximum common structure or placement fragment) was the same as the ratio of starting fragments for the whole of the XPoise follow-up library. The compounds were ensured to have sufficient coverage of chemical space by using a 2D similarity cut-off to ensure selected compounds were dissimilar to each other. This was checked by looking at the distributions of the inter-compound 2D similarity (Figure 5.12).

Subset 3.1 looks to achieve the same characteristics as subset 2 but was selected using the SQUONK diversity picker to minimise the Interaction Score whilst maximising the 2D diversity of the compounds. This achieved a similar coverage of the interaction space to subset 2 (Figure 5.18). Subset 3.2 consists of compounds selected by maximising the Interaction Score (No Frag). This score only includes interactions formed by the part of the compound that does not form the maximum common substructure with the fragment hits. Compounds are selected that are calculated to form novel interactions and should allow exploration of novel areas of interaction space. To maintain molecular diversity, the compounds were selected using the SQUONK diversity picker to maximise the Interaction Score (No Frag) whilst maximising 2D diversity. The diversity of interactions formed by the placement fragment that are already in the interaction grid was also found to be sufficient with a similar coverage to subset 2 and subset 3.1 (Figure 5.19).

Thus, both subset 3.1 and subset 3.2 were proposed for synthesis. Synthesis and subsequent screening of compounds from subset 3.1 should indicate whether the Interaction Score can be successfully used to prioritise compounds that are active by prioritising compounds that form interaction patterns similar to fragment hits. Compounds from subset 3 should indicate whether the Interaction Score can be used to prioritise compounds that interact with the target in novel ways. Due to a change in the priorities of targets for follow-up the subset was not synthesised, however it is hoped that this will be revisited, and the compounds will be synthesised in the future.

# 5.3 NUDT7

## *5.3.1 Introduction*

Human peroxisomal coenzyme A diphosphatase NUDT7 (hereafter referred to as NUDT7) is another member of the NUDIX family whose structure was determined by the SGC Oxford (PDB Code 5t3p; UniProt Code P0C024; *Srikannathasan et al.*, 2016). This was the first structure solved for this target. NUDT7 is of medical interest as its overexpression in mice led to a neurodegenerative disease linked to reduced CoA levels (Shumar *et al.*, 2015). A fragment screen was performed on the target and 21 high-confidence fragment hits were determined by X-ray crystallography. Each hit is shown in Figure 5.20.

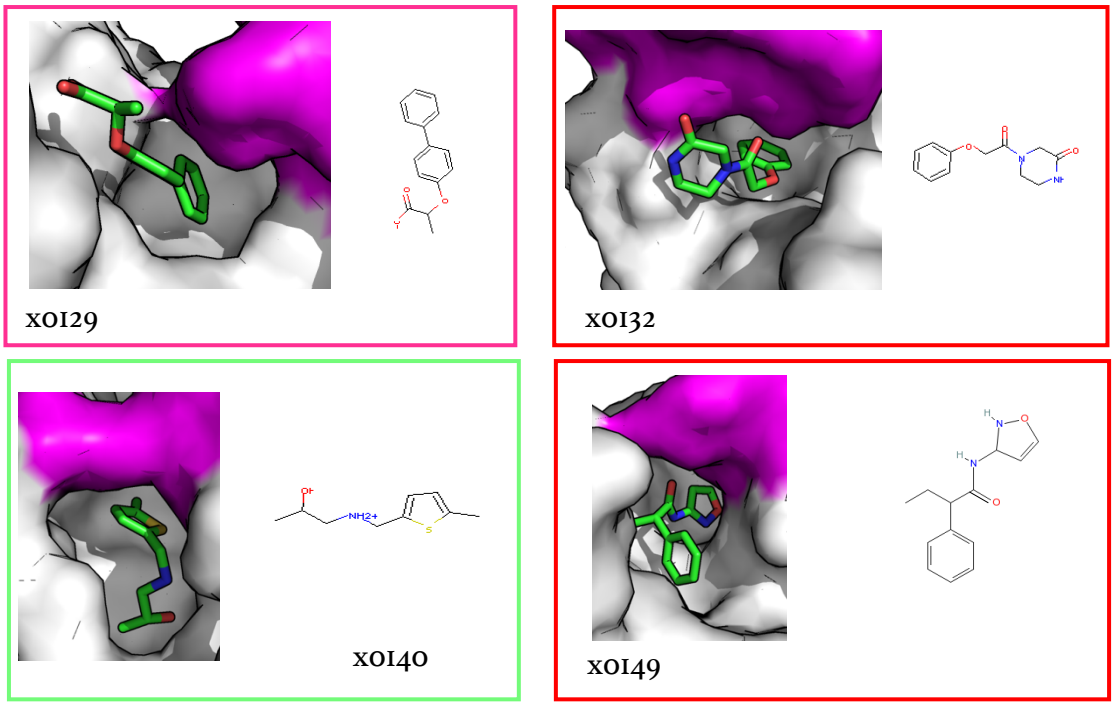



Figure 5.20 (continued on next page)

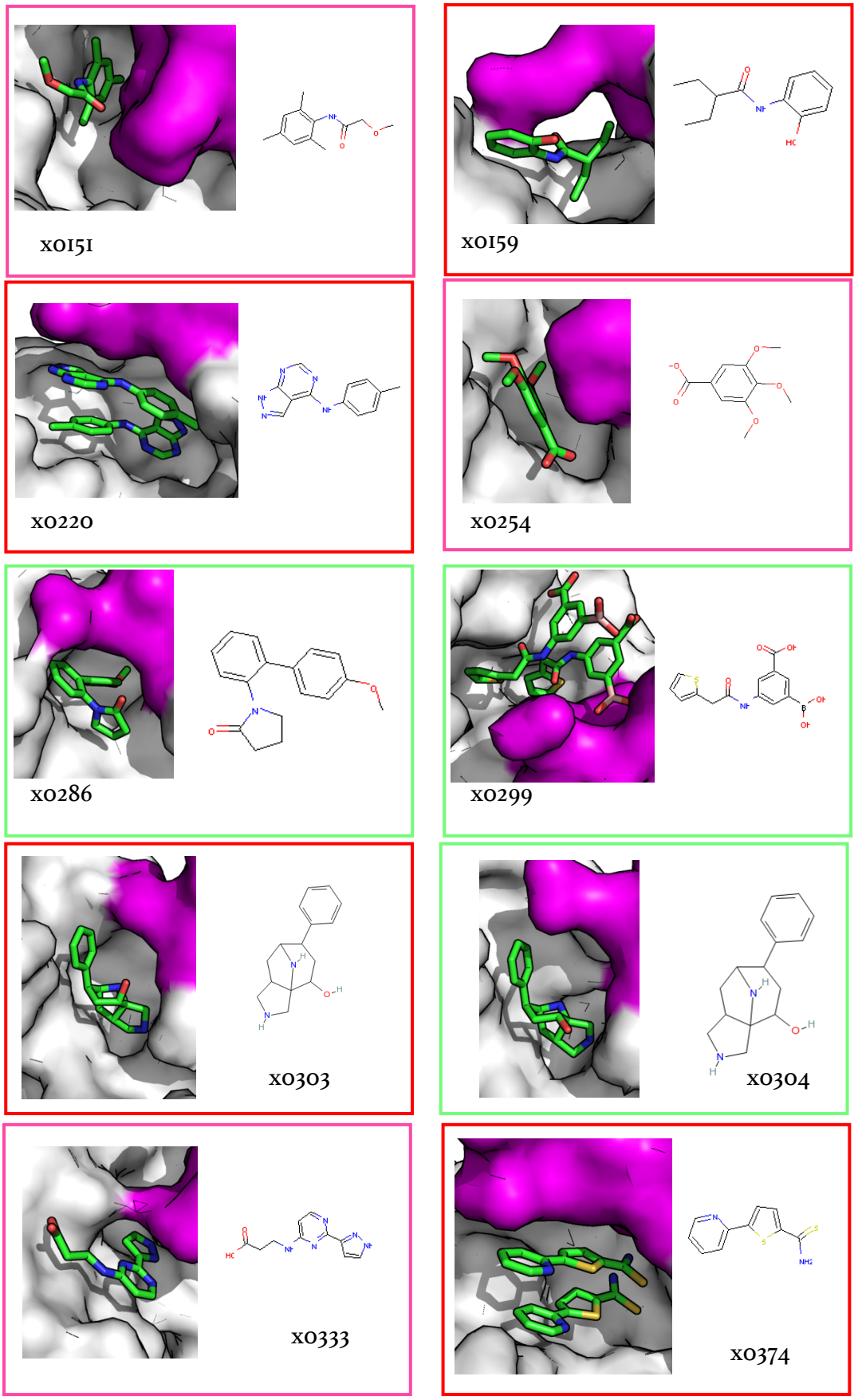



Figure 5.20 (continued on next page)

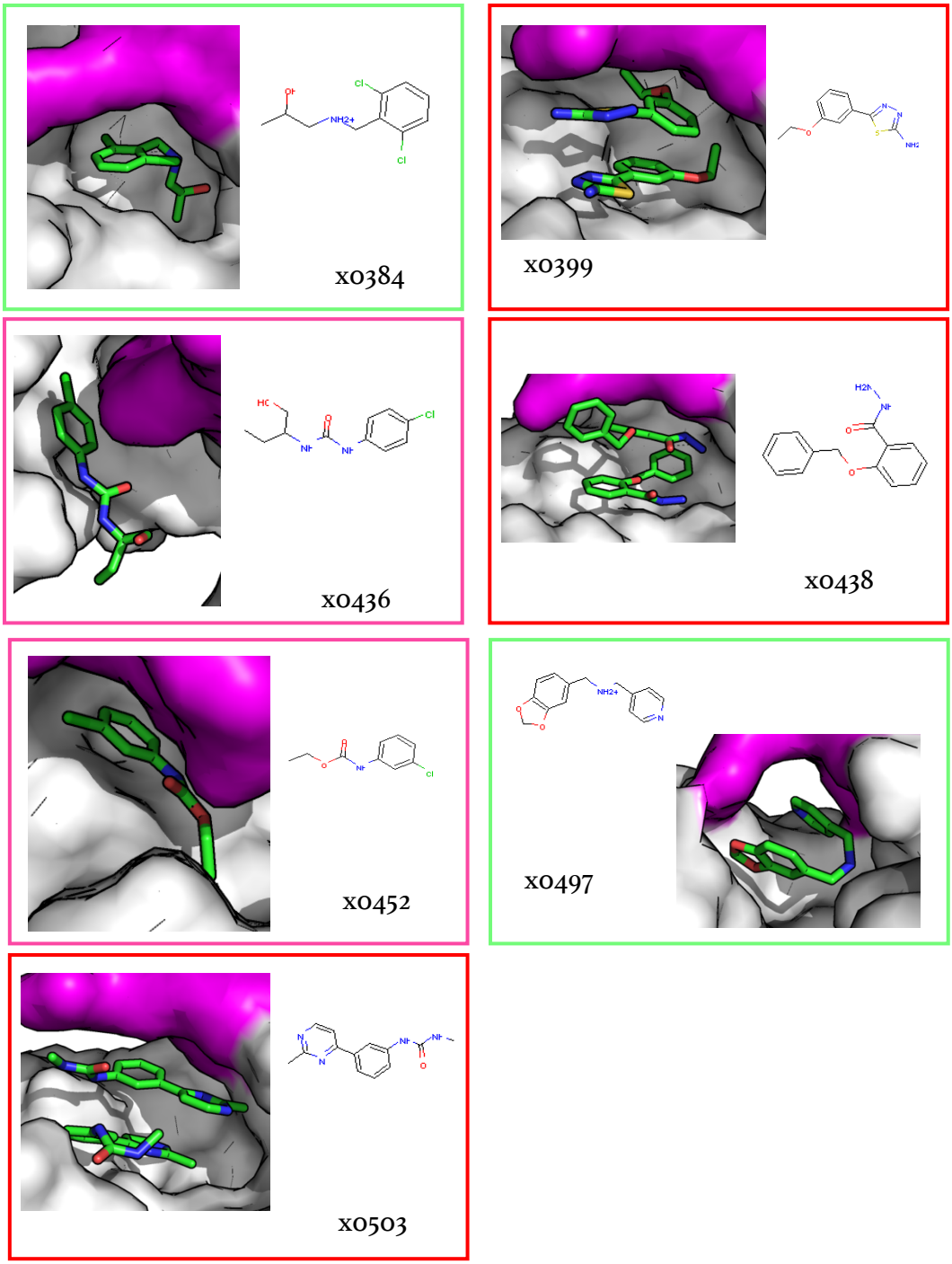



Figure 5.20 Structures and 2D depictions of each of the high confidence hits for target NUDT7. For each hit the code of the structure refers to an individual crystal structure from the SGC which can be found in Table 5.4. The coloured box outlines whether the fragment was excluded from the CRANkS grid. A red box indicates that the fragment hit clashes with the x0497 protein conformation. A pink box indicates that the x0497 hit clashes with the protein conformation. A green box indicates that the hit was used to construct the CRANkS grids. The pink part of the protein is the flexible loop discussed below.

One of the fragments (x0497) shown in Figure 5.20 occupied a bent conformation pointing further into the pocket. The follow-up team agreed that the fragment x0497 would be iterated looking to grow further into the area of the binding site shown in pink in Figure 5.21 Four of the hits formed interactions at the start of the pink area in the direction of growth – x0299, x0151, x0132 and x0432. A selection procedure was proposed to select compounds that could allow growth into that area of the binding pocket.

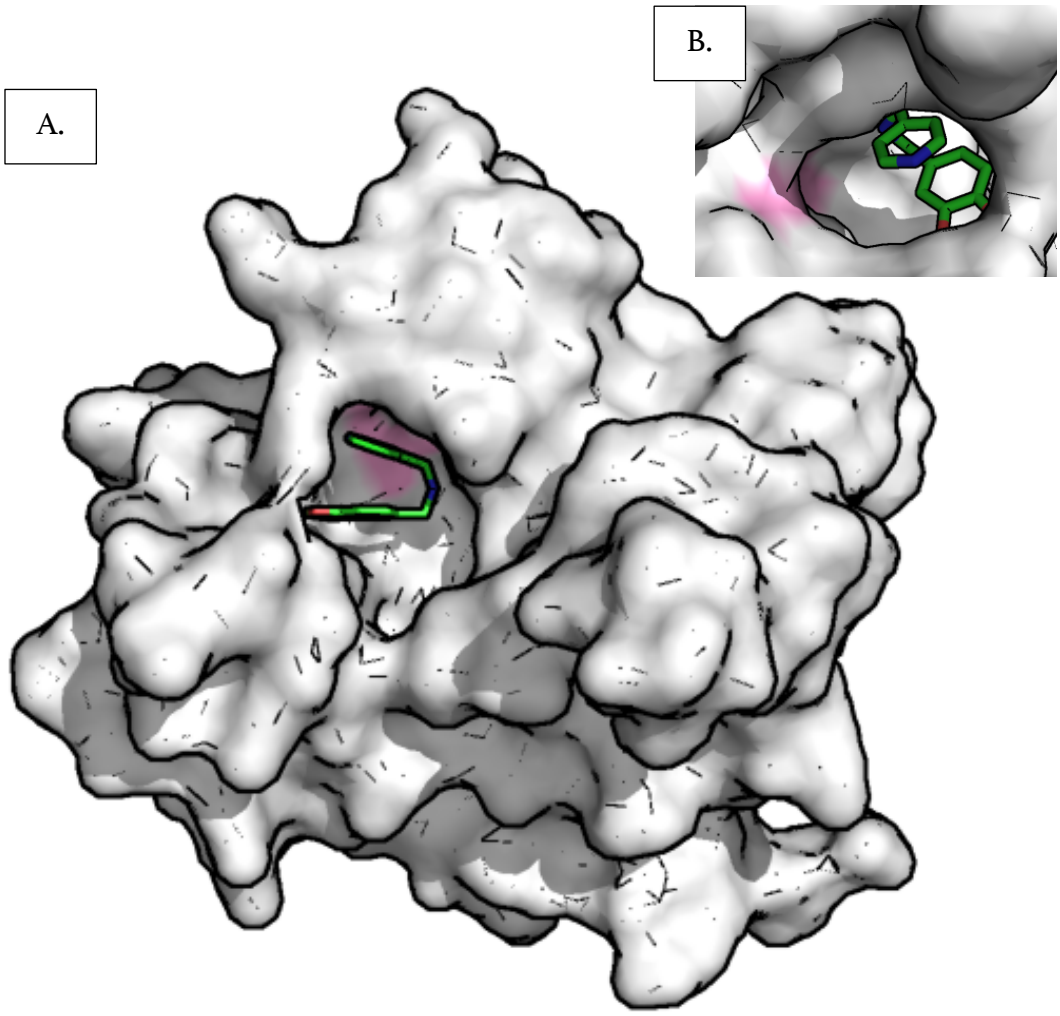



Figure 5.21. NUDT7 structure x0497. A shows the whole protein and ligand with the pink area indicating the area in which the fragment is to be grown into. B shows the structure rotated and zoomed to show more clearly the pink area highlighting the direction of growth of the fragment.

### *5.3.2 Method*

Follow-up compounds and conformers for those compounds were generated by Susan Leung using Reaction Vectors to iterate compounds and constrained docking to dock the compounds (*private communication)*. The CRANkS algorithm was therefore adapted to use conformers generated by an alternative method. The conformer found closest to the starting fragment x0497 by RMSD was taken forward for each compound.

Between each fragment hit there are conformational changes in the protein structure of NUDT7 (residue number 61 to 70 in PDB structure 5t3p). One loop is flexible and its side chain orientations change considerably between protein structures. This loop is shown in both "cartoon', "surface" and stick representations in Figures 5.22 to 5.24 and coloured pink in Figure 5.20. In the CRANkS grids it is useful to include any interactions within the area that I would like to extend into. However, interactions where the protein structure is significantly different would add noise to the system and was found to be detrimental to the performance of CRANkS in Chapter 3. The high confidence hits are shown in Table 5.4 and the protein-ligand structures used to construct CRANkS grids are coloured in green. Structures were rejected for two reasons. The first is if the protein structure clashed with the x0497 fragment hit with no room for growth of the molecule. This is shown in Figure 5.22. The second is if the ligand clashed with the x0497 protein conformation and this is shown in Figure 5.23. An example of a protein structure used in the CRANkS grid is shown in Figure 5.24.

| Protein-Ligand Structure | Used in the CRANkS Grids? | Reason for Rejection |
|---|---|---|
| x0129 | No | Protein clashes with x0497 hit and no room for growth |
| x0132 | No | Ligand clashes with x0497 protein conformation |
| x0140 | Yes | N/A |
| x0149 | No | Ligand clashes with x0497 protein conformation |
| x0151 | No | Protein clashes with x0497 hit and no room for growth |
| x0159 | No | Ligand clashes with x0497 protein conformation |
| x0220 | No | Ligand clashes with x0497 protein conformation |
| x0254 | No | Protein clashes with x0497 hit and no room for growth |
| x0286 | Yes | N/A |
| x0299 | Yes | N/A |
| x0303 | No | Ligand clashes with x0497 protein conformation |
| x0304 | Yes | N/A |
| x0333 | No | Protein clashes with x0497 hit and no room for growth |
| x0374 | No | Ligand clashes with x0497 protein conformation |
| x0384 | Yes | N/A |
| x0399 | No | Ligand clashes with x0497 protein conformation |
| x0436 | No | Protein clashes with x0497 hit and no room for growth |
| x0438 | No | Ligand clashes with x0497 protein conformation |
| x0452 | No | Protein clashes with x0497 hit and no room for growth |
| x0497 | Yes | N/A |
| x0503 | No | Ligand clashes with x0497 protein conformation |

Table 5.4. High-confidence protein-ligand structures of fragment hits with target NUDT7. Structures coloured in green were used in the CRANkS grids to score follow-up compounds. Structures coloured in red or pink were not used in the grids and the reasons for rejection are shown.

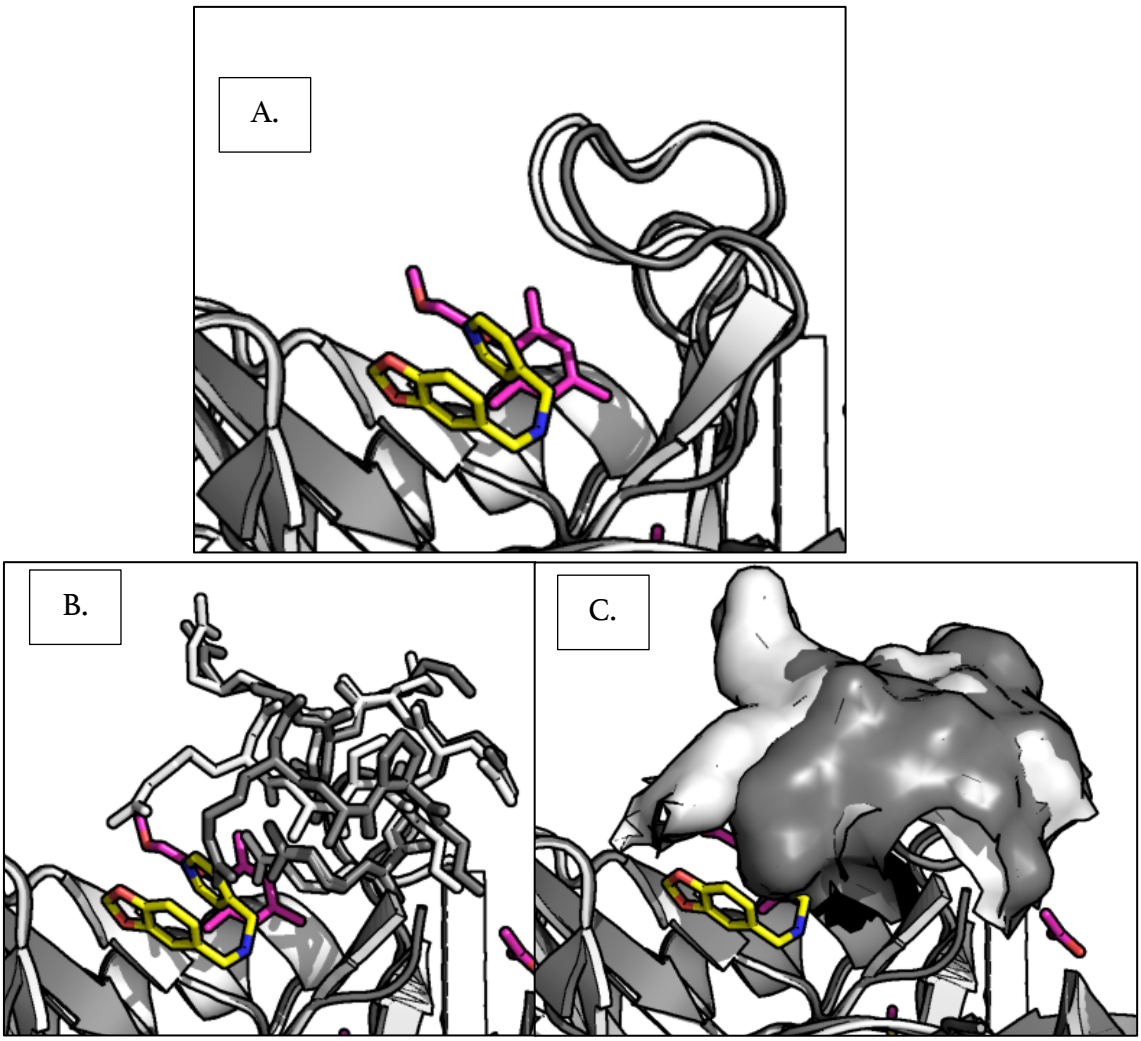



Figure 5.22 Structure x0151 shown with structure x0497. The ligand for x0497 is shown in yellow and the ligand for x0151 is shown in pink. The protein structure of x0497 is shown in white and the protein structure of x0151 is shown in grey. The flexible loop is shown in cartoon (A), sticks (B) and surface (C) representations. The x0497 ligand clashes with the protein structure of x0151. Structure x0151 is therefore excluded from the set of structures to build the CRANkS grids – the clash means that the compounds which were docked using constrained conformer generation using ligand x0497 will also likely clash with the structure, and the interactions formed in the x0151 will not be representative of what is possible in the x0497 environment.

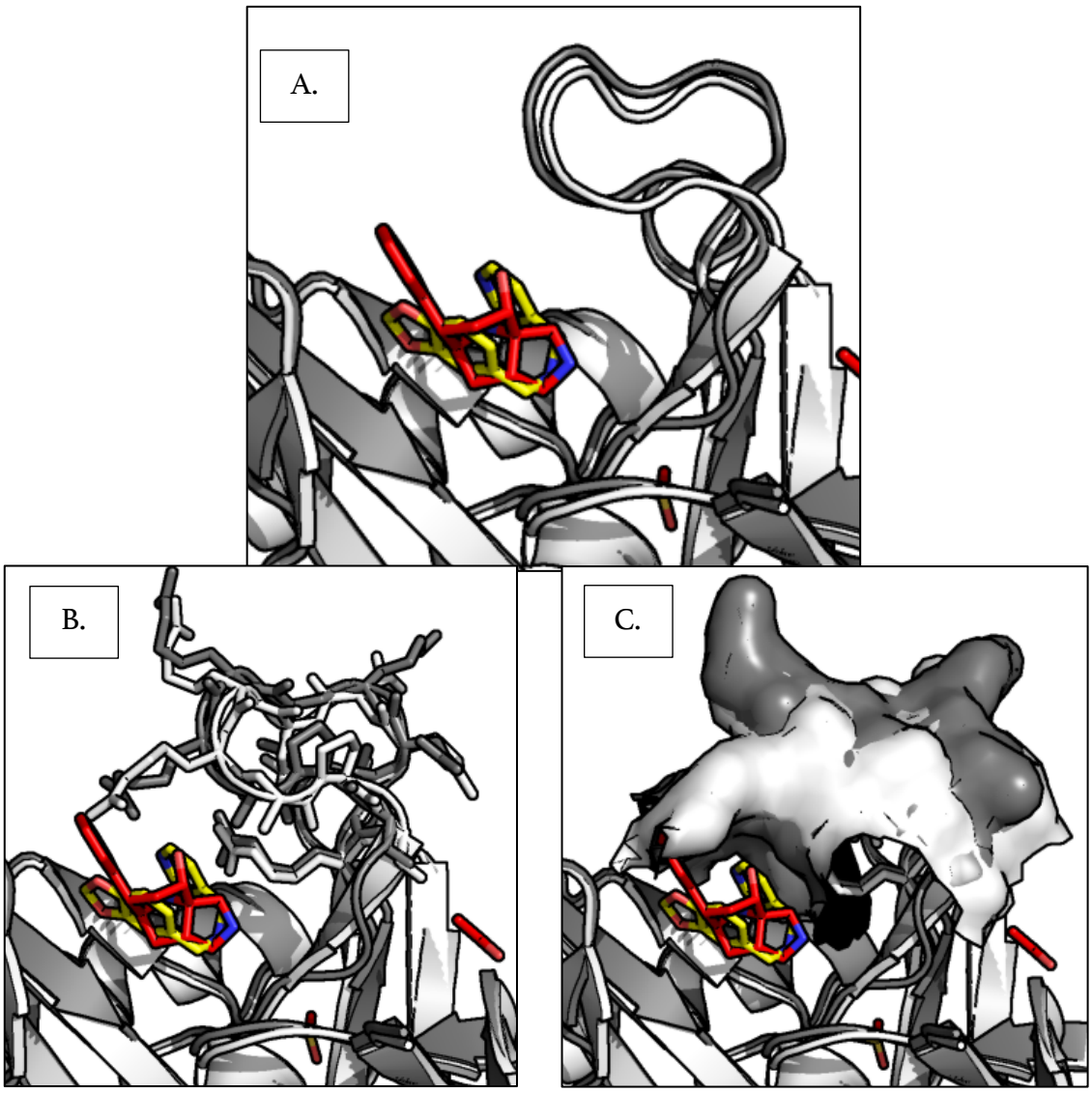



Figure 5.23 Structure x0303 shown with structure x0497. The ligand for x0497 is shown in yellow and the ligand for x0303 is shown in red. The protein structure of x0497 is shown in white and the protein structure of x0151 is shown in grey. The flexible loop is shown in cartoon (A), sticks (B) and surface (C) representations. The x0303 ligand clashes with the protein structure of x0497. Structure x0303 is therefore excluded from the set of structures to build the CRANkS grids – the clash means that the interactions formed in the x0303 will not be representative of what is possible in the x0497 environment.

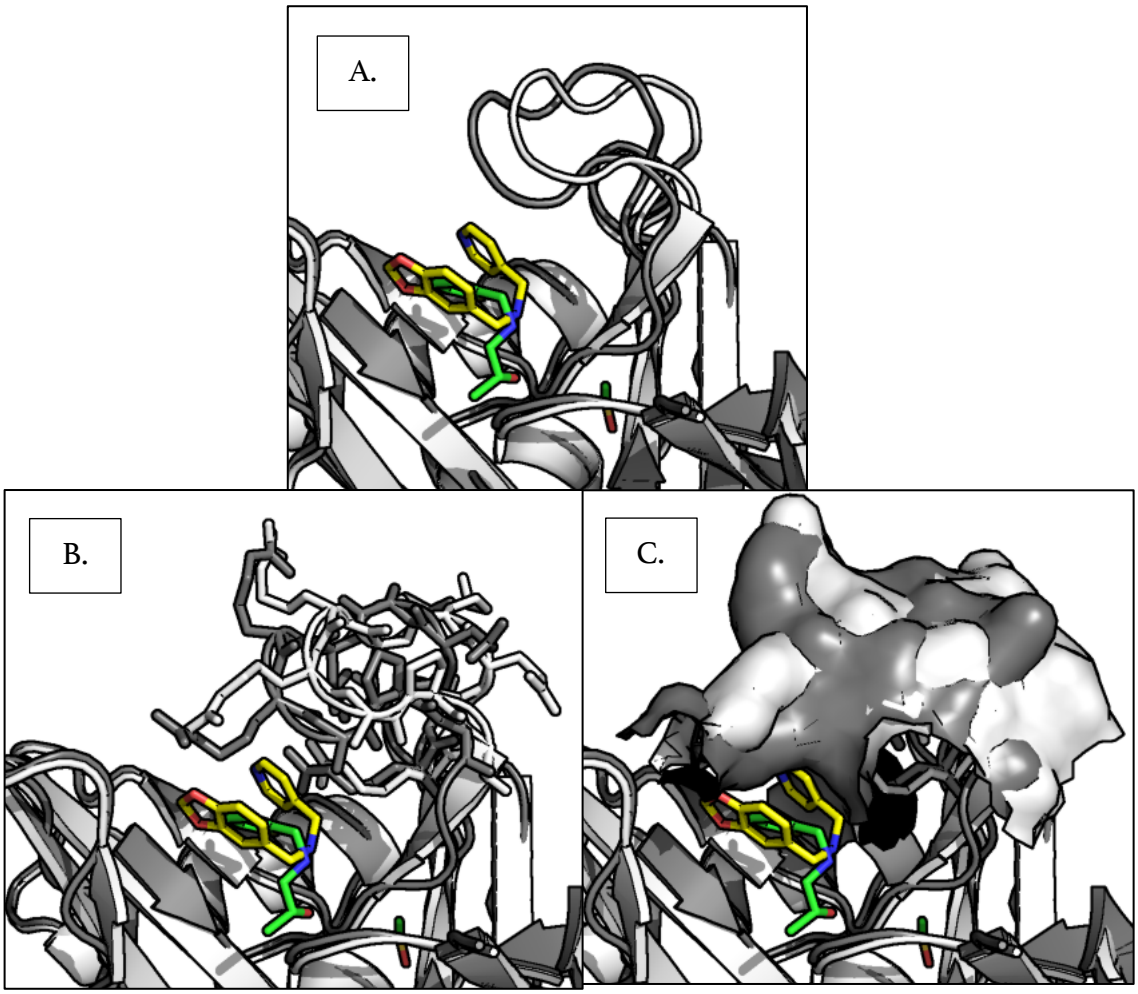



Figure 5.24 Structure x0140 with structure x0497. The ligand for x0497 is shown in yellow and the ligand for x0140 is shown in green. The protein structure of x0497 is shown in white and the protein structure of x0140 is shown in grey. The flexible loop is shown in cartoon (A), sticks (B) and surface (C) representations. Although the loops look different in all three representations, there are no clashes between either ligand and the protein, and the area for direction of growth has not been cut off (C). Therefore, this structure was one of the structures used to construct the CRANkS grids.

### *5.3.3 NUDT7 Results*

Each of the 607 candidate compounds was scored by CRANkS using the generated CRANkS grids. This yielded an Interaction Score for each compound. On inspection of the generated conformers, compounds with a low Interaction Score were not found to form novel interactions. An example of this is shown in Figure 5.25 for compound PDT 583. The predicted binding mode is shown in cyan and the

crystallographic structure of hit x0497 is shown in magenta. The compound was calculated to have an Interaction Score of 0.40, meaning the interactions formed overlaps highly with interactions of the interaction grid. The interaction grid is shown by transparent cylinders and the interactions formed by the compound are shown by opaque cylinders. The interaction pattern of the compound is the same as for the original hit x0497 with no novel interactions and this is why the compound has a low Interaction Score.

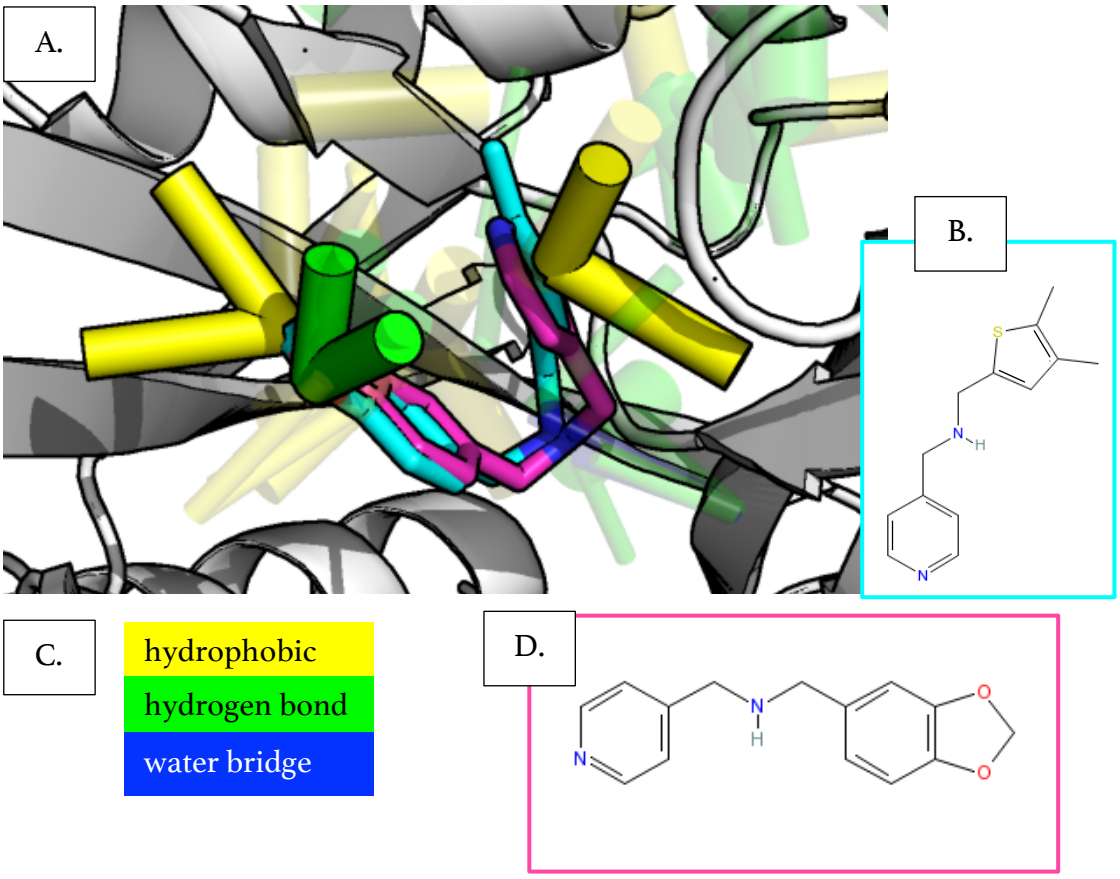



Figure 5.25. Compound PDT 583 is shown in cyan and the fragment hit from x0497 is shown in magenta (A). A shows the interaction grid by transparent cylinder where the width of the cylinder corresponds to the count of the cylinder. The interactions formed by compound PDT are shown by opaque cylinders. The compound was calculated to have an Interaction Score of 0.40 and is calculated to have an identical interaction pattern to the fragment hit, which can been seen as the transparent cylinders all overlap with translucent cylinders. B shows the 2D structure of PDT 583 and D shows the 2D structure of the fragment hit from x0497. C shows the key for the interaction types.

An example of a compound with a high Interaction Score (an Interaction Score of 0.91) is shown in Figure 5.26. The compound PDT 154 is predicted to form two new water bridge interactions and the water molecules forming these interactions are shown as red spheres. The compound also forms a novel hydrogen bond. The conformer is extending into the site with additional interactions, indicating that it could then be subjected to follow-up to explore the site even further.

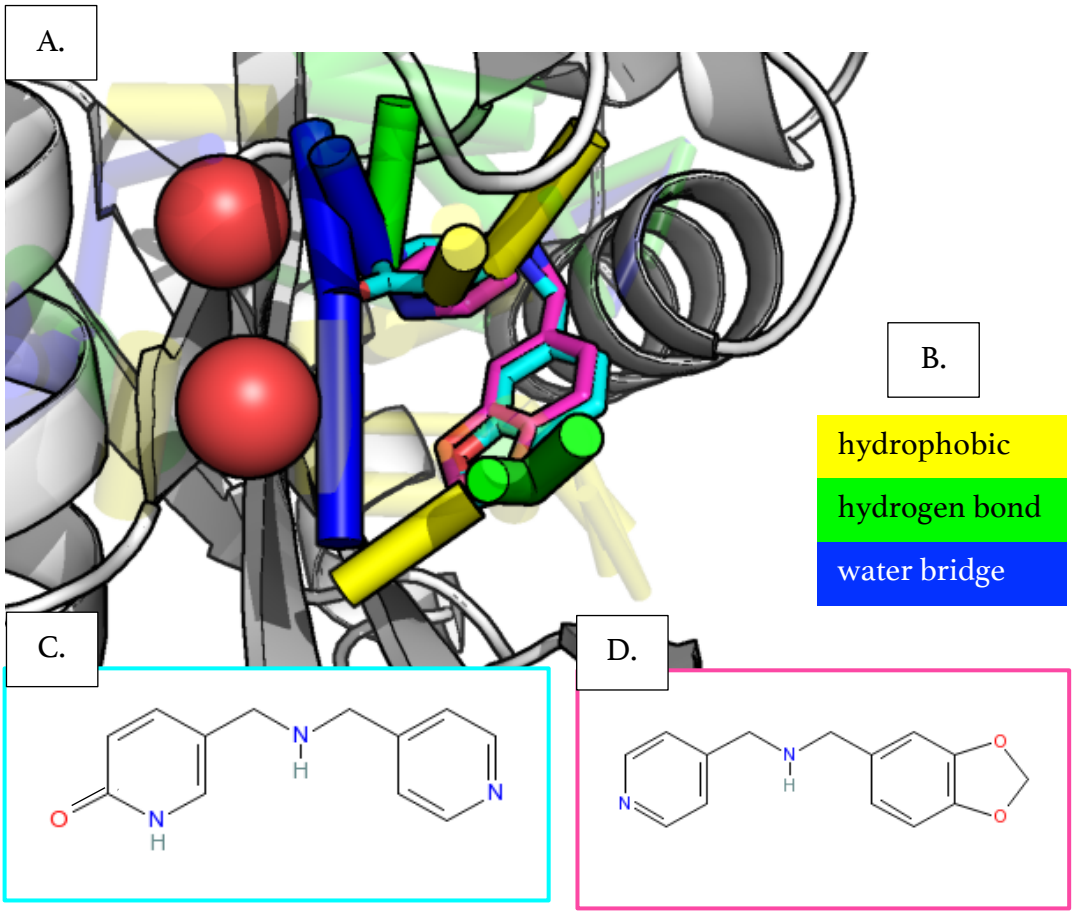
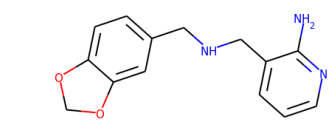


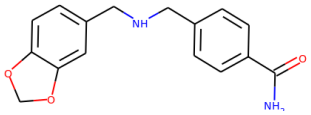

Figure 5.26. Compound PDT 154 is shown in cyan and the fragment hit from x0497 is shown in magenta (A). The interaction grid is shown by transparent cylinder where the width of the cylinder corresponds to the count of the cylinder. The interactions formed by compound PDT 154 bare shown by opaque cylinders. The compound was calculated to have an Interaction Score of 0.91 and is calculated to form novel water bridges, hydrogen bonds and hydrophobic interactions. These are the cylinders that do not overlap with translucent cylinders of the same colour. The water molecules involved in the water bridges are shown as red spheres. B shows the key for the interaction types in A. C and D show the 2D structure of PDT 154 and the fragment hit from x0497 respectively.

Given these trends, I decided it would be best to select compounds with a high Interaction Score. For this target, as the conformers were generated by constrained docking, all compounds adopt similar binding poses – a part of the molecule is always found in the same position and orientation. Thus, the Interaction Score in this case is a measure of novel interactions: molecules with a low Interaction Score form identical or very similar interactions to the x0497 ligand while molecules with a high Interaction Score are likely to form those previously observed interactions but also form additional novel interactions. To select compounds based on maximising the Interaction Score but maintain molecular diversity of compounds, the SQUONK diversity picker was used to select compounds for follow-up.

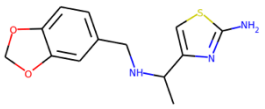
To investigate the potential of using the SQUONK diversity picker, a subset of 10 molecules was selected from the set of follow-up compounds. This size was selected because a number of avenues were being explored for follow-up,  so there was a limited budget for follow-up compounds that could be selected for synthesis using the CRANkS algorithm. The 10 compounds selected are shown in Figure 5.27.



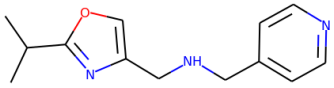
PDT85
Interaction Score = 1.0



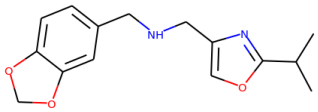
PDT83
Interaction Score = 1.0



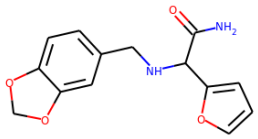
PDT371
Interaction Score = 1.0



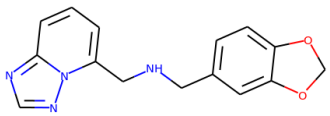
PDT498
Interaction Score = 0.84



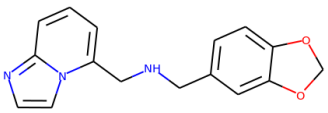
PDT331
Interaction Score = 0.84



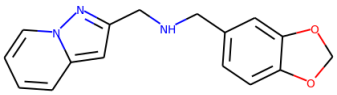
PDT10
Interaction Score = 1.0



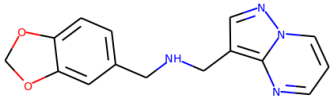
PDT5
Interaction Score = 0.80



PDT2
Interaction Score = 0.88



PDT512
Interaction Score = 0.89



PDT514
Interaction Score = 0.83

Figure 5.27. Subset of 10 compounds selected by the SQUONK diversity picker to maximise the Interaction Score. The Interaction Score is shown below each compound along with the compound name. Four compounds for which the interactions are shown in Figure 5.28 are shown within the pink box. All molecules are very similar with 6-membered and 5-membered fused aromatic rings containing nitrogen atoms.

By visual inspection the molecules do not look diverse: there are lots of similarities between the compounds. For example, the four compounds circled in the pink box in Figure 5.27 all share a fused set of aromatic rings consisting of a 5-membered and 6-membered ring containing nitrogen atoms. This is linked to the shared substructure of the fragment hit from x0497. To investigate this further, the binding pattern of the four compounds are shown in Figure 5.28.

It is clear that although the molecules may look similar in 2D, the binding pattern of all four are very different. Molecule PDT 5 forms hydrophobic and hydrogen bonds of the original hit shown at the bottom picture (Figure 5.28 A). The other molecules also form these interactions, however the interaction patterns of the conserved part of the molecule are all slightly different due to changes in orientation. The unique parts of the molecules, at the top of each pose, start to extend into other parts of the binding site near the interactions of other hits (Figure 5.28). PDT5 forms two novel water bridge interactions, as well as a novel hydrogen bond and hydrophobic interactions (Figure 5.28 A). PDT 2 forms a hydrogen bond and a π-π bond that was not formed by any of the fragment hits (Figure 5.28 B). PDT 512 and PDT 514 both form novel hydrophobic interactions and a novel hydrogen bond (Figure 5.28 C and D). The new interactions formed by each follow-up are unique compared to the interaction grid but also unique when compared to each other. This gives confidence that the follow-up compounds are sufficiently diverse in terms of interaction space.

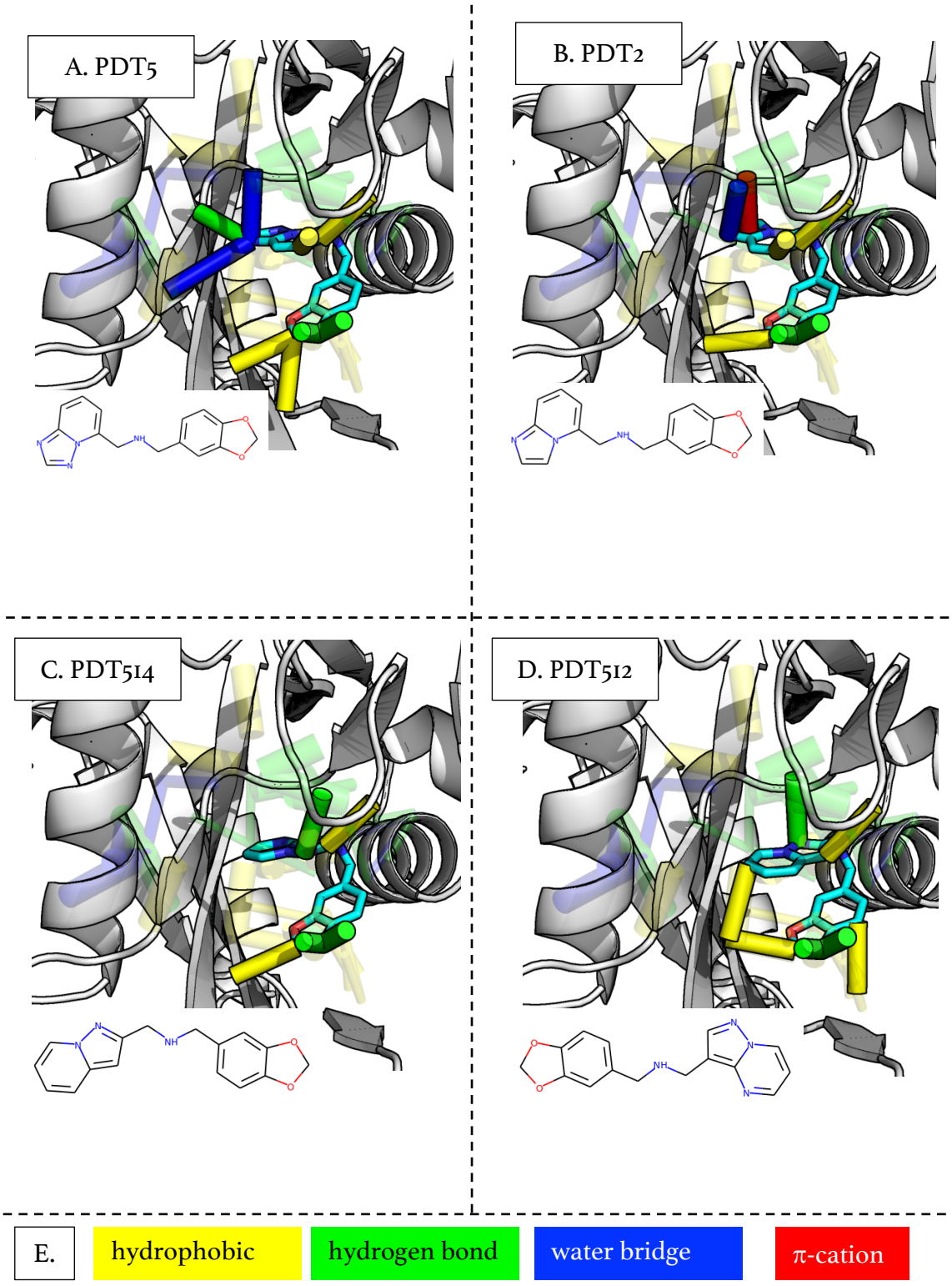



Figure 5.28 The conformations of four compounds selected by the SQUONK diversity picker for a subset of 10 compounds: PDT5 (A), PDT2 (B), PDT 514 (C) and PDT 515 (D). The interaction grid is shown by transparent cylinders where the width of the cylinder indicates the count of the interaction counter. The interactions formed by each compound are shown by opaque cylinders. Despite 2D molecular similarity the interaction patterns formed by each compound are different giving confidence in the diversity of interactions in the compound set. E shows the key for the colours of the interaction types.

### *5.3.4 NUDT7 Conclusions*

A subset of 10 compounds was proposed for follow-up for target NUDT7. The compounds and conformers were taken from a follow-up library generated by Susan Leung using Reaction Vectors and constrained docking (*private communication*). Therefore the CRANkS algorithm was modified to use conformers generated from an alternative source, rather than generating the conformers as part of the algorithm. This could allow other conformer generation tools such as docking to provide more accurate conformations for scoring by the algorithm in the future to avoid issues with conformer generation highlighted in Section 5.2.

Compounds were selected to form novel interactions whilst maintaining molecular diversity using the SQUONK diversity picker. Compounds were selected to maximise the Interaction Score and maintain 2D dissimilarity. The resulting selection was found to interact differently with the protein and form novel interactions. Thus, the subset was proposed for synthesis. Synthesis of the compounds would help to determine whether the CRANkS algorithm can be used to select follow-up compounds that interact with targets in novel ways compared to existing structural data for that target. Subsequent to proposal of purchase and synthesis of these follow-up compounds, the follow-up team decided to select compounds by visual inspection and synthetic intuition.

## 5.4 Conclusions

To summarise follow-up compounds were proposed for synthesis for two targets from the NUDIX protein family: NUDT22 and NUDT7. Compounds were selected in two ways. One subset for NUDT22 was selected for activity which I hypothesised could be chosen by minimising the Interaction Score. A second subset for NUDT22, and the subset for NUDT7 chose compounds selected to form novel interactions with the target which I hypothesised could be chosen by maximising the Interaction Score or an alternative to the Interaction Score which excluded the maximum common substructure between the candidate compound and the ligand from the score. Both subsets were also checked for coverage of the interaction space of the fragment hits and to have coverage of the molecular diversity of the set. Synthesis of the compounds would explore the potential of the CRANkS algorithm to select compounds for follow-up. Unfortunately, in both cases the follow-up compounds were not synthesised, however it is hoped this will be revisited in the future.

The process of selecting follow-up compounds using the CRANkS algorithm in a prospective setting highlighted two key issues with the algorithm. The first is the conformer generation part of the algorithm. This should be improved to be dependent on the number of rotatable bonds of the compound tested to improve efficiency. It should also cluster the conformers to reduce the inaccuracy of using conformers adopting the same conformation. This is partially addressed by combining the CRANkS grids with AutoDock to give AutoCRANkS as described in Chapter 4. The AutoCRANkS docking protocol should be combined with the

CRANkS algorithm to generate the conformers scored by CRANkS. The change in accuracy of the conformer generation should be determined.

The second issue identified is the lack of defined protocol for using the CRANkS scores to select compounds. This is addressed in Chapter 3 when a multi-objective optimisation method is described was found to select a diverse set of active scaffolds from a library of candidate compounds.

# Chapter 6

# Conclusions

In Chapter 1 I described the rising attrition rates and increasing cost of the stages of drug discovery. There is an emphasis on developing methods that will aid the selection of the optimal candidate compounds to take forward in the hit-to-lead and lead optimisation stages. This is coupled with an increase in structural data at the earliest stages of drug discovery. An example of this is the protocol of the i04-1 beamline at Diamond Light Source which can perform high-throughput fragment screening, screening 1000 compounds in one week. This yields a large amount of structural data of fragment hits bound to a target protein. There is currently no method that can make use of data consisting of multiple instances of hits binding to the same target to guide hit-to-lead optimisation in a systematic way. This sets the precedence for the work described in this thesis.

## 6.1 Chapter 2 Conclusions and Further Work

In Chapter 2, I describe the development of the CRANkS algorithm. This uses protein-ligand complexes of a target protein to build grids to describe the data. Novelty scores are generated for a set of candidate compounds based on how the compounds overlap with the CRANkS grids. Initial testing showed the algorithm was promising by exploring scaffold-hopping in HIV-1 protease datasets. Additionally the performance of the CRANkS Interaction Score at discriminating between actives and

inactives achieved a higher AUC than 2D fingerprint and PLIF methods for a BRD1 dataset.

Further work from this Chapter was to expand the molecular interactions calculated as part of the algorithm, in particular to include water-mediated hydrogen bonds. This is discussed in Chapter 3. Additionally, the conformer generation part of the algorithm, using the 3D Matched Molecular Pairs, was found to generate conformers of up to 20 Å away from the crystallographic structure of a ligand. Thus, further work would be required to update the algorithm to more accurate methods. Modifications to the conformer generation part of the algorithm are made in Chapter 3. In Chapter 4 the algorithm is combined with AutoDock which has a more rigorously tested conformer generation protocol.

Finally, a larger scale testing of the CRANkS algorithm on multiple targets was required to confirm the ability to of the algorithm to discriminate between active and inactive compounds and to prioritise novel scaffolds. This work was carried out in Chapter 3.

## 6.2 Chapter 3 Conclusions and Further Work

In Chapter 3 I describe the testing of the CRANkS algorithm on a number of diverse target datasets. The algorithm was tested in terms of active versus inactive discrimination and the prioritisation of a diverse range of scaffolds. The performance of the algorithm was target-dependent but was found to have a similar performance

to a number of commercial docking tools. The CRANkS scores were also found to prioritise more unique scaffolds than commonly-used fingerprinting methods.

A multi-objective optimisation method was developed to select compounds from a library given these results. This maximised the Element Score to prioritise novel compounds but minimised the Interaction Score to select compounds that were calculated to form similar interactions to ligands from protein-ligand complexes for the target of interest. The Pareto optimal solutions for the optimisation of these variables were selected as candidate compounds for datasets tested. An enrichment of unique scaffolds and unique active scaffolds were found in the selection, as well as a target-dependent enrichment of activity.

Further work on the algorithm would involve investigation on how protein-ligand structures should be selected to construct the CRANkS grids. This was discussed in the Chapter, but more work could be done to confirm the dependency of the performance of CRANkS on the shape overlap of the ligands used and the flexibility of the protein. This could involve clustering protein-ligand structures by the RMSD of the protein and running the algorithm on the clusters. This should be repeated using the shape similarity and protrusion of the ligands.

## 6.3 Chapter 4 Conclusions and Further Work

In Chapter 4, I describe “active-guided” docking. This combines the structural data from protein-ligand complexes for a target of interest with docking. This was

achieved by combining the CRANkS grids with AutoGrid maps used by AutoDock – AutoCRANkS. I have shown that by using structural knowledge about bound ligands it is possible to improve the ability of AutoDock to discriminate between actives and decoys.

Further work on this protocol would involve a much larger testing of AutoCRANkS on the remaining targets of the DEKOIS 2.0 dataset (Bauer *et al.,* 2013). This will determine the performance of AutoCRANkS on a much wider range of targets. AutoCRANkS could also be tested on the DUD-E dataset used in Chapter 3, to allow comparison to a number of other commercial docking tools and structure-based methods.

The use of AutoCRANkS in the multi-objective optimisation method should also be explored. The AutoCRANkS docking score generated for a candidate compound should be minimised to prioritise compounds that are active. The Element Score could be maximised to prioritise molecular novelty, or the Interaction Score could be maximised to prioritise the selection of compounds calculated to form novel interactions.

Finally, the CRANkS algorithm could be updated based on the methodology used to combine the CRANkS grids with the AutoGrid maps. The algorithm would be modified to use a finer grid spacing and a Gaussian distribution centred on each atom to construct the grid. The algorithm could also be updated to make use of the conformers generated for each compound using AutoCRANkS. The effect of these modifications on the algorithm should be explored.

## 6.4 Chapter 5 Conclusions and Further Work

In Chapter 5, I describe using the CRANkS algorithm in a prospective setting to select compounds for follow-up based on the results of fragment screening for two targets of interest. Compounds were selected based on two hypotheses. The first is that by minimising the Interaction Score compounds that are calculated to form similar interactions to the hits are selected, and that these are more likely to be active. The second is that by maximising the Interaction Score compounds are selected that form novel interactions compared to the hits, allowing exploration of new interaction space. In both cases the follow-up compounds have not yet been synthesised, but I hope this will be revisited in the future.

The process of selecting compounds as part of this work highlighting two key issues with the CRANkS algorithm. The first was the conformer generation step, which was found to be inefficient as it did not include any relation to the number of rotatable bonds. It was also likely to generate a false representation due to a lack of clustering of the generated conformations. Consequently, the algorithm must be modified to use a more accurate conformer generation tool. This is partially addressed in Chapter 4 by combining CRANkS grids with AutoDock. However, the CRANkS algorithm should also be modified to make use of the AutoCRANkS conformer generation.

The second issue emphasised by the selection of compounds for follow-up in Chapter 5 was a lack of a robust protocol to use the CRANkS scores to select compounds. This was addressed in Chapter 3 when a multi-objective optimisation process was derived to select compounds from a library of candidates. This was shown to select a diverse range of active scaffolds from a library of candidates for a number of targets. Further work should test this protocol in a prospective setting.

## Concluding Remarks

In this thesis I have described the development of algorithms to aid the selection of compounds from a library of candidates, during hit-to-lead optimisation, by using structural data of the target in question. The CRANkS algorithm was shown to select diverse active scaffolds using multi-objective optimisation. Combining protein-ligand structures with AutoGrid maps used by AutoDock improved the docking performance of AutoDock for a selection of targets. Overall, these tools will aid the process of small molecule design by using the structural data available for a target to guide the selection of compounds in hit-to-lead optimisation.

# References


Abergel C., *Molecular replacement: tricks and treats.,* Acta Crystallogr. Sect. D: Biol. Crystallogr., 2013 69, 2167-2173.

Anderson A. C., *The process of structure-based drug design.,* Chem. Biol., 2003, 10, 787-797.

Anighoro A., and Bajorath J., *Three-Dimensional Similarity in Molecular Docking: Prioritizing Ligand Poses on the Basis of Experimental Binding Modes.,* J. Chem. Inf. Model., 2016, 56, 580-587.

Armstrong M. S., Morris G. M., Finn P. W., Sharma R., Moretti L., Cooper R. I. and Richards W. G., *ElectroShape: fast molecular similarity calculations incorporating shape, chirality and electrostatics.,* J. Comput. Aided Mol. Des., 2010, 24, 789-801.

Ashton M., Barnhard J., Casset F., Charlton M., Downs G., Gorse D., Holliday J., Lahana R., and Willett P., *Identification of Diverse Database Subsets using Property-Based and Fragment-Based Molecular Descriptions,* Quant. Struct.-Act. Relat., 2002, 21, 598-604.

Badger J., *Crystallographic fragment screening*, Methods Mol. Biol., 2012, 841, 161-177.

Baell J. B., and Holloway G. A., *New Substructure Filters for Removal of Pan Assay Interference Compounds (PAINS) from Screening Libraries and for Their Exclusion in Bioassays.,* J. Med. Chem., 2010, 53, 2719-2740.

Bajusz D., Racz, A and Heberger K., *Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations?.,* J. Chem. Inf., 2015, 7, 20.

Ballester P. J. and Richards W. G., *Ultrafast shape recognition to search compound databases for similar molecular shapes.,* J. Comput. Chem., 2007, 28, 1711-1723.

Baringhaus K-H., and Matter H., *Cheminformatics in Drug Discovery* (Oprea T., ed.), 2004, 333-379, Wiley-VCH.

Bauer M. R., Ibrahim T. M., Vogel S. M., and Boeckler F. M., *Evaluation and optimization of virtual screening workflows with DEKOIS 2.0 - a public library of challenging docking benchmark sets.,* J. Chem. Inf. Model., 2013, 53, 1447-1462.

Bemis G. W. and Murcko M. A., *The properties of known drugs .1. Molecular frameworks.,* J. Med. Chem., 1996, 39, 2887-2893.

Bender A., *How similar are those molecules after all? Use two descriptors and you will have three different answers.,* Expert Opin.Drug Discov., 2010, 5, 1141-1151.

Bento A. P. Gaulton A., Hesey A., Bellis L. J., Chambers J., Davies M., Krüger F. A., Light Y., Mak L., McGlinchey S., Nowotka M., Papdatos G., Santos R., and Overington J. P., *The ChEMBL bioactivity database: an update.,* Nuc. Acids Res., 2014, 42, 1083-1090.

Berman H. M., Westbrook J., Feng Z., Gilliland G., Bhat T. N., Weissig H., Shindyalov I. N., and Bourne P. E., *The Protein Data Bank.,* Nuc. Acids Res., 2000, 28, 235-242.

Bessman M. J. Frick D. N., and O'Handley S. F., *The MutT proteins or "Nudix" hydrolases, a family of versatile, widely distributed, "housecleaning" enzymes.,* J. Biol Chem., 1996, 11, 25059-25062.

Biasini M., *PV.,* https://zenodo.org/record/20980#.W6aA3i-ZNsM, 2015, pv: v1.8.1, Available at https://github.com/biasmv/pv. Accessed on 1 January 2016.

Bibette J., *Gaining confidence in high-throughput screening.,* Proc. Acad. of Sciences of U.S.A., 2012, 109, 649-650.

Bissantz C., Kuhn B., and Stahl M., *A Medicinal Chemist's Guide to Molecular Interactions.,* J. Med. Chem., 2010, 53, 5061-5084.

Bleicher K. A., Böhm H. J., Müller K., Alanine A. I., *Hit and lead generation: Beyond high-throughput screening.,* Nat. Rev. Drug. Discov., 2003, 5, 368-378.

Bodnarchuk M. S., *Water, water, everywhere...It's time to stop and think.,* Drug Discov. Today, 2016, 21, 1139-1146.

Bohacek R.S., McMartin C. and Guida, W.C., *The art and practice of structure-based drug design: A molecular modelling perspective.,* Med. Res. Rev., 1996,16, 3-50.

Bohm H-J., Flohr A. and Stahl M., *Fluorine in medicinal chemistry.,* Drug Discov. Today. Technol., 2004, 1, 217-224.

Bradley A. R., Wall I. D., Green D. V. S., Deane C. M and Marsden B. D., *OOMMPPAA: A Tool To Aid Directed Synthesis by the Combined Analysis of Activity and Structural Data.,* J. Chem. Inf. Model., 2014, 54, 2636-2646.

Bradley E. K., Miller J. L., Saiah E., and Grootenhuis P. D. J., *Informative library design as an efficient strategy to identify and optimize leads: Application to cyclin-dependent kinase 2 antagonists.,* J. Med. Chem., 2003, 46, 4360-4364.

Bradley A. R. Wall I. D., Green D. V. S., Deane C. M and Marsden B. D., University of Oxford, 2015, *Development of Tools to Provide Prioritisation and Guidance in the Development of Chemical Probes and Small Molecule Leads*.

Brik A., and Wong C. *HIV-1 protease: mechanism and drug discovery,* Org. Biomol., Chem., 2003, 1, 5-14.

Brooijmans N., *Inside the Mind of a Medicinal Chemist: The Role of Human Bias in Compound Prioritization during Drug Discovery.,* PLOS One, 2012, 7, e48476

Brown A. C. And Fraser T. R., *On the connection between chemical constitution and physiological action.,* J. Anat. Physiol., 1868, 2, 224-242.

Brown D. G., Gagnon M.M., and Bostrom J., *Understanding Our Love Affair with p-Chlorophenyl: Present Day Implications from Historical Biases of Reagent Selection.,* J. Med. Chem., 2015., 58, 2390-2405.

Brown N., Royal Society of Chemistry, 2016, *In Silico Medicinal Chemistry Computational Methods to Support Drug Design*.

Brumbaugh J., Di Stefano B., Wang X., Borkent M., Forouzmand E., Clowers K. J., Ji F., Schwarz B. A., Kalocsay M., Elledge S. J., Chen Y., Sadreyev R. I., Gygi S. P., Hu G., Shi Y., and Hochedlinger K., *Nudt21 Controls Cell Fate by Connecting Alternative Polyadenylation to Chromatin Signaling.,* Cell., 2018, 172, 106-120.

Brünger A. T., *Free R value: cross-validation in crystallography.,* Methods Enzymol., 1997, 277, 366-396.

Chadwick A. T., and Segall M. D., *Overcoming psychological barriers to good discovery decisions.,* Drug Discov. Today, 2010, 15, 561-569.

Chaput L., Martinez-Sanz J., Saettel N., and Mouawad L., J., *Benchmark of four popular virtual screening programs: construction of the active/decoy dataset remains a major determinant of measured performance.,* Cheminformatics, 2016, 8, 56.

Marvin Beans 16.8.1.0, 2016, ChemAxon (http://www.chemaxon.com) Available at https://www.chemaxon.com/download/marvin-suite/ Accessed on 1 June 2016.

Cherkasov A., Muratov E. N., Fourches D., Varnek A., Baskin I. I., Cronin M., Dearden J., Gramatica P., Martin Y. C., Todeschini R., Consonni V., Kuz'min V. E., Cramer R., Benigni R., Yang C., Rathman J., Terfloth L., Gasteiger J., Richard A. and Tropsha A., *QSAR Modeling: Where Have You Been? Where Are You Going To?,* J. Med. Chem., 2014, 57, 4977-5010.

Christopher J. A., Brown J., Doré A. S., Errey J. C., Koglin M., Marshall F. H., Myszka D. G., Rich R. L., Tate C. G., Tehan B., Warne T., and Congreve M., *Biophysical fragment screening of the β1-adrenergic receptor: identification of high affinity arylpiperazine leads using structure-based drug design*., J. Med. Chem., 2013, 56, 3446-3455.

Ciociola A. A., Cohen L. B., & Kulkarni P., *How Drugs are Developed and Approved by the FDA: Current Process and Future Directions.,* Am. J. Gastroenterol., 2014, 109, 620–623.

Coman R. M., Robbins A. H., Goodenow M. M., Dunn B. M. and McKenna R., *High-resolution structure of unbound human immunodeficiency virus 1 subtype C protease: implications of flap dynamics and drug resistance.,* Acta Cryst., 2008, 64, 754-763.

Congreve M., Carr R., Murray C., and Jhoti H., *A 'rule of three' for fragment-based lead discovery?,* Drug Discov. Today, 2003, 8, 876-877.

Cook D., Brown D., Alexander R., March R., Morgan P., Satterthwaite G., and Pangalos M. N., *Lessons learned from the fate of AstraZeneca's drug pipeline: a five-dimensional framework.,* Nat. Rev. Drug Discov., 2014, 13, 419–431.

Cox O. B., Von Delft F., and Brennan P., University of Oxford, 2016, *Design and utilisation of a poised fragment library against epigenetic proteins.*

Cox O. B., Krojer T., Collins P., Monterio O., Talon R., Bradley A., Fedorov O., Amin J., Marsden B. D., Spencer J., von Delft F. and Brennan P. E., *A poised fragment library enables rapid synthetic expansion yielding the first reported inhibitors of PHIP(2), an atypical bromodomain.,* Chem. Sci., 2016, 7, 2322-2330.

Cresset, Forge, Activity Atlas, Available at: http://www.cresset-group.com/activity-atlas/

Cumming J. G., Davis A. M., Muresan S., Haeberlein M., and Chen H., *Chemical predictive modelling to improve compound quality.,* Nat. Rev. Drug Discov., 2013, 12, 948-962.

Da C. and Kireev D., *Structural Protein-Ligand Interaction Fingerprints (SPLIF) for Structure-Based Virtual Screening: Method and Benchmark Study.,* J. Chem. Inf. Model., 2014, 54, 2555-2561.

Da Costa D., Agathanggelou A., Perry T., Weston V., Petermann E., Zlatanou A., Oldreive C., Wei W., Stewart G., Longman J., Smith E., Kearns P., Knapp S., and Stankovic T., *BET inhibition as a single or combined therapeutic approach in primary paediatric B-precursor acute lymphoblastic leukaemia.,* Blood Cancer Journal, 2013, 3, e126.

Danziger D. J., and Dean P. M., *Automated Site-Directed Drug Design: A General Algorithm for Knowledge Acquisition about Hydrogen-Bonding Regions at Protein Surfaces,* Proc. R. Soc. Lond. B., 1989, 236, 101-113.

De Luca G., Ventura I., Sanghez V., Russo M, T., Ajmone-Cat M. A., Cacci E., Martire A., Popoli P., Falcone G., Michelini F., Crescenzi M., Degan P., Minghetti L., Bigami M., and Calamandrei G., *Prolonged lifespan with enhanced exploratory behavior in mice overexpressing the oxidized nucleoside triphosphatase hMTH1*, Aging Cell, 2013, 12, 695-705.

Dean P. M., Lloyd D. G., and Todorov N. P., *De novo drug design: Integration of structure-based and ligand-based methods.,* Curr. Opin. Drug. Discov. Devel., 2004, 7, 347-353.

Deb K., Pratap A., Agarwal S. and Meyarivan T., *A fast and elitist multiobjective genetic algorithm: NSGA-II.,* IEEE Trans. Evo. Comput., 2002, 6, 182-197.

Django (Version 1.9) [Computer Software]. (2015). Retrieved from *https://djangoproject.com.* Accessed on 1 September 2015.

Dobson C.M., *Chemical space and biology.,* Nature 2004, 432,824-828.

Dolinsky T. J., Czodrowski P., Li H., Nielsen J. E., Jensen J. H., Klebe G., Baker N. A., *PDB2PQR: expanding and upgrading automated preparation of biomolecular structures for molecular simulations.,* Nuc. Acid Res., 2007, 35, W22-W25.

Dolinsky, T. J., Nielsen J. E., McCammon J. A., and Baker N. A., *PDB2PQR: an automated pipeline for the setup of Poisson-Boltzmann electrostatics calculations.,* Nuc. Acid Res., 2004, 32, W665-W667.

Duan J., Dixon, S. L., Lowrie J. F. and Sherman W., *Analysis and comparison of 2D fingerprints: Insights into database screening performance using eight fingerprint methods.,* J. Mol. Graph. Model., 2010, 29, 157-170.

Ebejer J-P., Morris G. M. and Deane C. M., *Freely Available Conformer Generation Methods: How Good Are They?,* J. Chem. Inf. Model., 2012, 52, 1146-1158.

Ehmki E. S. R., and Rarey M., *Exploring Structure–Activity Relationships with Three-DimensionalMatched Molecular Pairs—A Review,* Chem. Med. Chem., 2018, 13, 482-489.

Erickson J. A., Jalaie M., Robertson D. H., Lewis R. A., Vieth M., *Lessons in molecular recognition: the effects of ligand and protein flexibility on molecular docking accuracy.,* J. Med. Chem., 2004, 47, 45-55.

Erlanson D. A., Fesik S. W., Hubbard R. E., Jahnke W., and Jhoti H., *Twenty years on: the impact of fragments on drug discovery.,* Nat. Rev. Drug Disc., 2016, 15, 605-619.

Ewing T. J., Makino S., Skillman A. G. and Kuntz I. D., *DOCK 4.0: Search strategies for automated molecular docking of flexible molecule databases.,* J. Comput. Aided Mol. Des., 2001, 15, 411-428.

Filippakopoulos P., Qi J., Picaud S., Shen Y., Smith W. B., Fedorov O., Morse E. M., Keates T., Hickman T. T., Felletar I., Philpott M., Munro S., McKeown M. R., Wang Y., Christie A. L., West N., Cameron M. J., Schwartz B., Heightman T. D., La Thangue N., French C. A., Wiest O., Kung A. L., Knapp S., and Bradner J. E., *Selective inhibition of BET bromodomains.,* Nature, 2010, 468, 1067-1073.

Fiskus W., Sharma S., Qi J., Valenta J. A., Schaub L. J., Shah B., Devaraj S., G., T., Santhana G. T., Leveque C., Portier B. P., Iyer S., Bradner J. E., and Bhalla K. N., *BET Protein Antagonist JQ1 Is Synergistically Lethal with FLT3 Tyrosine Kinase Inhibitor (TKI) and Overcomes Resistance to FLT3-TKI in AML Cells Expressing FLT-ITD.,* Molecular Cancer Therapeutics, 2014, 13, 1142–54.

Fortin F., De Rainville F.-M., Gardner M.-A., Parizeau M., Gagne C., *DEAP: A Python Framework for Evolutionary Algorithms.,* Journ. Machine Learn. Res., 2012, 13, 2171-2175.

Friesner R. A., Banks J. L., Murphy R. B., Halgren T. A., Klicic J. J., Mainz D. T., Repasky M. P., Knoll E. H., Shelley M., Perry J. K., Shaw D. E., Francis P., Shenkin P. S., *Glide: A new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy.,* J. Med. Chem., 2004, 47, 1739-1749.

Gaieb Z., Liu S., Gathiaka S., Chiu M., Yang H. W., Shah C. H., Feher V. A., Walters W. P., Kuhn B., Rudolph M. G., Burley S. K., Gilson M. K., and Amaro R. E., *D3R Grand Challenge 2: blind prediction of protein-ligand poses, affinity rankings, and relative binding free energies.,* J. Comp. Aid. Mol. Des., 2018, 32, 1-20.

Galloway W.R., Isidro-Llobet A., Spring D. R., *Diversity-oriented synthesis as a tool for the discovery of novel biologically active small molecules.,* Nat. Commun., 2010, 1, 80.

Gedeck P., Rohde B., Bartels C., *QSAR - How good is it in practice? Comparison of descriptor sets on an unbiased cross section of corporate data sets.,* J. Chem. Inf. Model., 2006, 46, 1924-1936.

Gillet V.J., *Diversity selection algorithms.,* Wiley Interdiscip. Rev.: Comput. Mol. Sci., 2011, 1, 580-589.

Gilson M. K., Liu T., Baitaluk M., Nicola G., Hwang L., Chong J., *BindingDB in 2015: A public database for medicinal chemistry, computational chemistry and systems pharmacology.,* Nuc. Acid Res., 2016, 44, D1045-D1063.

Gobbi A., and Poppinger D., 1998, *Genetic optimization of combinatorial libraries.,* Biotechnology and Bioengineering, 61, 47-54.
Goodford P. J., J. Med. Chem., *A computational-procedure for determining energetically favorable binding-sites on biologically important macromolecules.,* 1985, 28, 849-857.

Grant J. A., Gallardo M. A., Pickup B. T., *A fast method of molecular shape comparison: A simple application of a Gaussian description of molecular shape.,* J. Comput. Chem., 1996, 17, 1653-1666.

Groom C. R., Bruno I. J., Lightfoot M. P. and Ward S. C., *The Cambridge Structural Database.,* Acta Cryst., 2016, 171-179.

Haka D., Platypus [Computer Software], Retrieved from https://github.com/Project-Platypus/Platypus.git

Halgren T. A., Murphy R. B., Friesner R. A., Beard H. S., Frye L. L., Pollard W. T., Banks J. L., *Glide: A new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy.,* J. Med. Chem., 2004, 47, 1750-1759.

Hann M. M., *Molecular obesity, potency and other addictions in drug discovery.,* Med. Chem. Commun., 2011, 2, 349-355.

Hartenfeller M., Eberle M., Meier P., Nieto-Oberhuber C., Altmann K.-H., Schneider G., Jacoby E., and Renner S., *A Collection of Robust Organic Synthesis Reactions for In Silico Molecule Design.,* J. Chem. Inf. Model., 2011, 51, 3093-3098

Hartenfeller M., Zettl H., Walter M., Rupp M., Reisen F., Proschak E., Weggen S., Stark H., and Schneider G., *DOGS: Reaction-Driven de novo Design of Bioactive Compounds.,* PLOS Comput. Biol., 2012, 8, e1002380.

Hawkins D. M., *The Problem of Overfitting.,* J. Chem. Inf. Comp. Sci., 2004, 41, 663-670.

Hopkins A.L., *Network pharmacology: the next paradigm in drug discovery.,* Nat. Chem. Bio., 2008, 4, 682-690.

Hu Y., Stumpfe D., Bajorath J., *Computational Exploration of Molecular Scaffolds in Medicinal Chemistry.,* J. Med. Chem., 2016, 59, 4062-4076.

Huang S-Y., Grinter S.Z. and Zou X., 2010, *Scoring functions and their evaluation methods for protein-ligand docking: recent advances and future directions.,* Phys. Chem. Chem. Phys, 12, 12899-12908.
Huang N., Shoichet B. K. and Irwin J. J., *Benchmarking sets for molecular docking.,* J. Med. Chem., 2006, 49. 6789-6801.

Hugunin J., *The Python Matrix Object: Extending Python for Numerical Computation.* Proceedings of the Third Python Workshop, Reston, Va., Dec. 1995; NumPy: a numerical extension for the computer language Python. Version 1.7. Available at http://www.python.org/topics/scicomp/numpy.html Accessed on 1 September 2015.

Hung, A. W., Silvestre, H. L., Wen, S., Ciulli, A., Blundell, T. L., and Abell, C., *Application of fragment growing and fragment linking to the discovery of inhibitors of Mycobacterium tuberculosis pantothenate synthetase.,* Angewandte Chemie., 2009, 48, 8452-8456.

Ibrahim T. M., Bauer M. R., and Boeckler F. M., *Applying DEKOIS 2.0 in structure-based virtual screening to probe the impact of preparation procedures and score normalization.,* J. Chem. Inf., 2015, 7.

Ichihara O., Barker J., Law R. J., Whittaker M., *Compound Design by Fragment-Linking.,* Mol. Inf., 2011, 30, 298-306.

IUCR, R factor from *http://reference.iucr.org/dictionary/R_factor*, retrieved September 6, 2015.

Jain A. N., *Surflex: fully automatic flexible molecular docking using a molecular similarity-based search engine.,* J. Med. Chem., 2003, 46, 499-511.

Jain A. N., *Scoring functions for protein-ligand docking.*, Curr. Protein Pept. Sci., 2006, 7, 407-420.
Jasail S., Hu Ye., Vogt M., Bajorath J., *Activity-relevant similarity values for fingerprints and implications for similarity searching.,* Chem. Inf. Sci., 2015, 5, 591.
Johnson M. A., and Maggiora G. M., *Concepts and Applications of Molecular Similarity*., John Wiley, 1990.

Jones G., Willett P., Glen R. C., Leach A. R., and Taylor R., *Development and validation of a genetic algorithm for flexible docking.,* J. Mol. Biol., 1997, 267, 727-748.

Jones G., Willett P., and Glen R. C., *Molecular recognition of receptor sites using a genetic algorithm with a description of desolvation.,* J. Mol. Biol., 245, 43-53, 1995.

Jones E., Oliphant E., and Peterson P., et al., SciPy: Open Source Scientific Tools for Python, 2001, Available at: *http://www.scipy.org/* Accessed on 1st September 2017.

Jubb H. C., Higueruelo A. P., Ochoa-Montano B., Pitt W. R., Ascher D. B. and Blundell T. L., *Arpeggio: A Web Server for Calculating and Visualising Interatomic Interactions in Protein Structures.,* J. Mol. Bio., 2017, 429, 365-371.

Khamis M. A. and Gomaa W., *Comparative assessment of machine-learning scoring functions on PDBbind 2013.,* Engineering Applications of Artificial Intelligence, 2015, 45, 136.

Kim S., Thiessen P. A., Bolton E. E., Chen J., Fu G., Gindulyte A., Han L., He J., He S., Shoemaker B. A., Wang J., Yu B., Zhang J. and Bryant S. H., *PubChem Substance and Compound databases.,* Nucleic Acids Res., 2016, 44, D1202-D1213.

Kitchen D. B., Decornez H., Furr J. R. and Bajorath J., *Docking and scoring in virtual screening for drug discovery: Methods and applications.,* Nat. Rev. Drug Discov., 2004, 3, 935-949.

Klebe G., *Virtual ligand screening: strategies, perspectives and limitations.,* Drug Discovery Today, 2006, 11, 580-594.

Knockaert M., Greengard P., and Meijer L., *Pharmacological inhibitors of cyclin-dependent kinases.,* Trends in Pharmacological Sciences, 2002, 23, 417-425.

Korb O., Finn P. W., and Jones G., *The cloud and other new computational methods to improve molecular modelling.,* Exp. Op. Drug Discov., 2014, 9, 1121-1131.

Koutsoukas A., Paricharak S., Galloway, W. R., Spring D. R., Ijzerman A. P., Glen R. C., Marcus D., and Bender A., *How Diverse Are Diversity Assessment Methods? A Comparative Analysis and Benchmarking of Molecular Descriptor Space.,* J. Chem. Inf. Model, 2014, 54, 230-242.

Kozakov D., Hall D. R., Jehle S., Luo L., Ochiana S. O., Jones E. V., Pollastri M., Allen K. N., Whitty A., and Vajda S., *Ligand deconstruction: Why some fragment binding positions are conserved and others are not.,* Proc. Natl. Acad. Sci. U. S. A., 2015, 112, E2585-2594.

Kutchukian P. S., Vasilyeva N. Y., Xu J., Lindvall M. K., Dillon M. P., Glick M., Coley J. D. and Labute P., J., *A widely applicable set of descriptors.,* Mol. Graphics Modell., 2000, 18, 464-477.

Lajiness M. S., Maggiora G. M., and Shanmugasundaram V., *Assessment of the consistency of medicinal chemists in reviewing sets of compounds.,* J. Med. Chem., 2004, 47, 4891-4896.

Landrum G., RDKit: Open-Source Cheminformatics. Retrieved from http://www.rdkit.org. Accessed on 1 September 2015.

Landrum G. A., Penzotti J. E., and Putta S., *Feature-map vectors: a new class of informative descriptors for computational drug discovery.,* J. Comput. Aided Mol. Des., 2006, 20, 751-762.
Lewell X.Q., Judd D.B., Watson S.P., and Hann M.M., *RECAP - Retrosynthetic combinatorial analysis procedure: A powerful new technique for identifying privileged molecular fragments with useful applications in combinatorial chemistry.,* J. Chem. Inf. Comput. Sci., 1998, 38, 511-522.

Lewis R. A., Mason J S. and McLay I. M., *Similarity measures for rational set selection and analysis of combinatorial libraries: The diverse property-derived (DPD) approach.,* J. Chem. Inf. Comput. Sci., 1997, 37, 599-614.

Lexa K. W., and Carlson H. A., *Protein flexibility in docking and surface mapping.,* Q. Rev. Biophys., 2012, 45, 301-343.

Li H., Robertson A. D., and Jensen J. H., *Very fast empirical prediction and rationalization of protein pKa values.,* Proteins, 2005, 61, 704-721.

Li Y., Dai J. H., Song M. G., Fitzgerald-Bocarsly P., and Kiledjian M., *Dcp2 Decapping Protein Modulates mRNA Stability of the Critical Interferon Regulatory Factor (IRF) IRF-7.,* Mol. Cell. Bio., 2012, 32, 1164-1172.

Lipinski C., and Hopkins A., *Navigating chemical space for biology and medicine.,* Nature, 2004, 432, 855-861.

Lipinski, C. A., Lombardo, F., Dominy, B. W., and Feeney, P. J., *Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings.,* Advanced Drug Deliv. Rev., 2001, 46, 3-26.

Liu Z., Li Y., Han L., Li J., Liu J., Zhao, Z., Nie W., Liu Y., and Wang R., *PDB-wide collection of binding data: current status of the PDBbind database.,* Bioinformatics, 2015, 31, 405-412.

Liu T., Lin Y., Wen X., Jorissen R. N. and Gilson M. K., *BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities.,* Nucleic Acids Res., 2007, 35, D198-D201.

Lusher S.J., McGuire R., Azevedo R., Boiten J.W., Van Schaik R.C. and de Vlieg, J., *A molecular informatics view on best practice in multi-parameter compound optimization.,* Drug Discov. Today, 16, 555-568.

Maggiora, G. M., *On outliers and activity cliffs - Why QSAR often disappoints.,* J. Chem. Inf. Model.,2006, 46, 1535–1535.

Maggiora G. M., Vogt M., Stumpfe D., and Bajorath J., *Molecular Similarity in Medicinal Chemistry.,* J. Med. Chem. 2014, 57, 3186–3204.

Manchester J., and Czerminski, R., *SAMFA: Simplifying molecular description for 3D-QSAR.,* J. Chem. Inf. Model., 2008, 48, 1167-1173.

Maneewongvatana S., and Mount D. M., Analysis of approximate nearest neighbor searching with clustered point sets., 1999, Retrieved from: https://arxiv.org/abs/cs/9901013

Martin Y. C., Willett P., Lajiness M., Johnson M., Maggiora G., Martin E., Bures M. G., Gasteiger J., Cramer R. D., Pearlman R. S., and Mason J.S., *Diverse viewpoints on computational aspects of molecular diversity.,* Journal of Combinatorial Chemistry, 2001, 3, 231-250.

McGregor M.J., Muskal S. M., *Pharmacophore fingerprinting. 1. Application to QSAR and focused library design.,* J. Chem. Inf. Comput. Sci., 1999, 39, 569-574.

McLennan A. G., *The Nudix hydrolase superfamily.,* Cell Mol. Life. Sci., 2006, 63, 123-143.

Medina-Franco J.L., Martinez-Mayorga K., and Meurice N., *Balancing novelty with confined chemical space in modern drug discovery.,* Expert Opin. Drug Discov., 2014, 9, 151-165.

Merkel D., *Docker: lightweight Linux containers for consistent development and deployment.,* Linux, 2014, 4, 2.

Mertz J. A., Conery A. R., Bryant B. M., Sandy P., Balasubramanian S., Mele D. A., Bergeron L., Balsubramanian S., Male D. A., Bergeron L., and Sims R. J., *Targeting MYC dependence in cancer by inhibiting BET bromodomains.,* Proceedings of the National Academy of Sciences of the United States of America, 2011, 108, 16669–74.

*Molecular Operating Environment (MOE)*, 2013.08; 2018, Chemical Computing Group ULC, 1010 Sherbooke St. West, Suite #910, Montreal, QC, Canada, H3A 2R7.

Moitessier N.. Englebienne P., Lee D., Lawandi J., and Corbeil C. R., *Towards the development of universal, fast and highly accurate docking/scoring methods: a long way to go.,* British Journal of Pharmacology, 2008, 153, S7-S26.

Morgan H. L., *The Generation of a Unique Machine Description for Chemical Structures-A Technique Developed at Chemical Abstracts Service.,* J. Chem. Doc., 1965, 5, 107-113.

Morgan S., Grootendorst P., Lexchin J., Cunningham C. and Greyson D., *The cost of drug development: A systematic review.,* Health Policy, 2011, 100, 4-17.

Morris G. M., Goodsell D. S., Halliday R. S., Huey R., Hart W. E., Belew R. K., Olson A. J., *Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function.,* J. Comp. Chem., 1998, 19, 1639-1662.

Morris G.M., Huey R., Lindstrom W., Sanner M. F., Belew R. K., Goodsell D. S., and Olson A. J., *AutoDock4 and AutoDockTools4: Automated Docking with Selective Receptor Flexibility.,* J. Comp. Chem., 2009, 30, 2785-2791.

Mysinger M. M., Carchia M., Irwin J. J., and Schoichet B. K., *Directory of Useful Decoys, Enhanced (DUD-E): Better Ligands and Decoys for Better Benchmarking.,* J. Med. Chem., 2012, 55, 6582-6594.

Nguyen K. T., Blum L. C., van Deursen R., and Reymond J. L., *Classification of Organic Molecules by Molecular Quantum Numbers.,* Chem. Med. Chem., 2009, 4, 1803-1805.

Nicolaou C. A. and Brown N., *Multi-objective optimisation methods in drug design,* Drug Discov. Today: Technol., 2013, 10, e427-e435.

Nicolotti O., Gillet V. J., Fleming P. J., and Green D. V. S., *Multiobjective Optimization in Quantitative Structure–Activity Relationships: Deriving Accurate and Interpretable QSARs.,* J. Med. Chem., 2002, 45, 5069-5080.

O'Boyle N. M., Banck M., James C. A., Morley C., Vandermeersch T., and Hutchison G. R., *Open Babel: An open chemical toolbox.,* J. Chem. Inf., 2011, 3, 33.

Page B. D. G., Valerie N. C. K., Wright R. H. G., Wallner O., Isaksson R., Carter M., Rudd S. G., Loseva O., Jemth A. S., Almof I., Font-Mateu J., Llona-Minguez S., Baranczewski P., Jeppsson F., Homan E., Almqvist H., Axelsson H., Regmi S., Gustavsson A. L., Lundback T., Scobie M., Stromberg K., Stenmark P., Beato M., and Helleday T., *Targeted NUDT5 inhibitors block hormone signaling in breast cancer cells.,* Nat. Comm., 2018, 9, 250.

Pareto, V., Cours d'économie politique : professé à l'Université de Lausanne. Pichon: Paris, 1896.

Patel D., Bauman J. D., and Arnold E., *Advantages of crystallographic fragment screening: Functional and mechanistic insights from a powerful platform for efficient drug discovery.,* Prog. Biophys. Mol. Biol., 2014, 116, 92-100;

Pearce N. M., Krojer T., Bradley A. R., Collins P., Radosław P. N., Talon R., Marsden B. D., Kelm S., Shi J., Deane C. M., and von Delft F., *A multi-crystal method for extracting obscured crystallographic states from conventionally uninterpretable electron density.,* Nat. Comms., 2017, 8, 15123.

*The PyMOL Molecular Graphics System*, Version 2.0 Schrödinger, LLC.

Radifar M., Yuniarti N. and Istyastono E. P., *PyPLIF: Python-based Protein-Ligand Interaction Fingerprinting.,* Bioinformation, 2013, 9, 325-328.

Rarey M., Kramer B., Lengauer T., and Klebe G., *A fast flexible docking method using an incremental construction algorithm.,* J. Mol. Biol., 1996, 261,470-489.

Reddy A. S., and Zhang S., *Polypharmacology: drug discovery for the future.,* Expert Rev. Clin. Pharmacol., 2013, 6, 41–47.

Reddy M. R., Reddy C. R., Rathore R. S., Erion M. D., Aparoy P., Reddy R. N., and Reddanna P., *Free Energy Calculations to Estimate Ligand-Binding Affinities in Structure-Based Drug Design.,* 2015, Curr. Pharm. Design., 20, 3323-3337.

Ren J., He Y., Chen W. Y., Chen T. T., Wang G., Wang Z., Xu Z. J., Luo X. M., Zhu W. L., Jiang H. L., Shen J. S. and Xu Y. C., *Thermodynamic and Structural Characterization of Halogen Bonding in Protein-Ligand Interactions: A Case Study of PDE5 and Its Inhibitors.,* 2014, J. Med. Chem., 57, 3588-3593.

Reymond J.-L., *The Chemical Space Project.,* Acc. Chem. Res., 2015, 48, 722-730.

Riniker S., and Landrum G. A., *Better Informed Distance Geometry: Using What We Know To Improve Conformation Generation.,* J. Chem. Inf. Model., 2015, 55, 2562-2574.

Ripphausen P., Stumpfe D., and Bajorath J., *Analysis of structure-based virtual screening studies and characterization of identified active compounds.,* Future Med. Chem., 2012, 4, 603-613.

Rogers D. and Hahn M., *Extended-Connectivity Fingerprints*., 2010, J. Chem. Inf. Model., 50, 742-754.

Rohrer S. G. and Baumann K., *Maximum Unbiased Validation (MUV) Data Sets for Virtual Screening Based on PubChem Bioactivity Data.,* J. Chem. Inf. Model., 2009, 49, 169-184.

Ross G. A., Morris G. M., and Biggin P. C., *One Size Does Not Fit All: The Limits of Structure-Based Models in Drug Discovery.,* J. Chem. Theory Comput., 2013, 9, 4266-4274.

Ross G. A., Morris G. M., and Biggin P. C., *Rapid and Accurate Prediction and Scoring of Water Molecules in Protein Binding Sites.,* PLOS ONE., 2012, 7, e32036.

Roth H.J., *There is no such thing as 'diversity"! Opinion.,* Curr. Opin. Chem. Biol., 2005, 9, 293-295.

Roughley S. D. and Jordan A.M., *The Medicinal Chemist's Toolbox: An Analysis of Reactions Used in the Pursuit of Drug Candidates.,* J. Med. Chem., 2011, 54, 3451-3479.

Roy A.S.A, Project FDA Report, 24 April, 2012, Available at: http://www.manhattan-institute.org/html/stifling-new-cures-true-cost-lengthy-clinical-drug-trials-6013.html. Accessed 1 August 2016.

Roy A., Srinivasan B., and Skolnick J., *PoLi: A Virtual Screening Pipeline Based on Template Pocket and Ligand Similarity.,* J. Chem. Inf. Model., 2015 55, 1757-1770.

Salentin S. Schreiber S., Haupt V. J., Adasme M. F., and Schroeder M., *PLIP: fully automated protein-ligand interaction profiler.,* Nucl. Acids. Res., 2015, 43, W443-W447.

Samaranayake G. J., Huynh M., and Rai P., *MTH1 as a Chemotherapeutic Target: The Elephant in the Room*, Cancers, 2017, 9, 47.

Schafer T., Kriege N., Humbeck L., Klein K., Koch O., Mutzel P., *Scaffold Hunter: a comprehensive visual analytics framework for drug discovery.,* 2017, J. Chem. Inf., 9, 1758-2946.

Schneider G., Neidhart W., Giller T., and Schmid G., *"Scaffold-hopping" by topological pharmacophore search: A contribution to virtual screening.,* Angewandte Chemie, 1999, 19, 2894-2896.
Schneider P., and Schneider G., *De Novo Design at the Edge of Chaos,,* J. Med. Chem., 2016, 12, 4077-4086.

Schneider G., and Fechner U., *Computer-based de novo design of drug-like molecules.,* Nat. Rev. Drug, Discov., 2005, 4, 649-663.

Schneider G., *Designing the molecular future.,* J. Comp.-Aided Mol. Des., 2012, 26, 115-120.

Schneider G., Wiley-VCH, 2013, *De Novo Molecular Design*.

Schombrug K., Ehrlich H. -C., Stierand K, and Rarey M., *From Structure Diagrams to Visual Chemical Patterns.,* J. Chem. Inf. Model., 2010, 50, 1529-1535.

*The PyMOL Molecular Graphics System*, Version 1.8 Schrödinger, LLC.

Schuffenhauer A., Ertl P., Roggo S., Wetzel S., Koch M. A., and Waldmann H., *The Scaffold Tree – Visualization of the Scaffold Universe by Hierarchical Scaffold Classification.,* J. Chem. Inf. Model., 2007, 47, 47-58.

Schwartz J., Awale M., and Reymond J. L., *SMIfp (SMILES fingerprint) Chemical Space for Virtual Screening and Visualization of Large Databases of Organic Molecules.,* J. Chem. Inf. Model., 2013, 53, 1979-1989.

Sheridan R. P., Singh S. B., Fluder E. M., and Kearsley S. K., *Protocols for bridging the peptide to nonpeptide gap in topological similarity searches.,* 2001, J. Chem. Inf. Comput. Sci., 41, 1395-1406.

Shoichet B. K., *Drug Discovery: nature's pieces.,* Nat. Chem., 2013, 5, 9-10.

Shultz M., *Improving the Plausibility of Success with Inefficient Metrics.,* Bioorg. Med. Chem. Lett., 2013, 23, 5980-5981.

Shumar S. A., Kerr E. W., Geldenhuys W. J., Montgomery G. E., Fagone P., Thirawatananond P., Saavedra H., Gabelli S. B. and Leonardi R., *Nudt19 is a renal CoA diphosphohydrolase with biochemical and regulatory properties that are distinct from the hepatic Nudt7 isoform.,* J. Bio. Chem., 2018, In press, Retrieved from http://www.jbc.org/.

Shumar S. A., Fagone P., Alfonso-Pecchio A., Gray J. T., Rehg J. E., Jackowski S., and Leonardi R., *Induction of Neuron-Specific Degradation of Coenzyme A Models Pantothenate Kinase-Associated Neurodegeneration by Reducing Motor Coordination in Mice,* 2015, PLOS ONE, 10, e0130013.

Sielecki T.M., Boylan J. F., Benfield P. A., and Trainor G. L., *Cyclin-dependent kinase inhibitors: Useful targets in cell cycle regulation.,* J. Med. Chem., 2000, 43, 1-18.

Silvestre H. L., Blundell T. L., Abell C., and Ciulli A., *Integrated biophysical approach to fragment screening and validation for fragment-based lead discovery*, Proc. Natl. Acad. Sci. U. S. A., 2013, 110, 12984-12989.

Snarey M., Terret N. K., Willett P. and Wilton D. J., *Comparison of algorithms for dissimilarity-based compound selection.,* J. Mol. Graph. Model., 1998, 15, 372-385.

Sousa S. F., Ribeiro A. J. M., Coimbra J. T. S., Neves R. P. P., Martins S. A., Moorthy N. S. H. N., Fernandes P. A., and Ramos M. J., *Protein-Ligand Docking in the New Millennium - A Retrospective of 10 Years in the Field.,* Curr. Med. Chem., 2013, 20, 2296-2314.

Spyrakis F., and Cavasotto C. N., *Open challenges in structure- based virtual screening: receptor modelling, target flexibility consideration and active site water molecules description.,* Arch. Biochem. Biophys., 2015, 583, 105-119.

Sridhar A., Ross G. A., and Biggin P. C., *Waterdock 2.0: Water placement prediction for Holo-structures with a pymol plugin.,* PLOS ONE, 2017, 12, e0172743.

Stumpfe, D., and Bajorath, J., *Exploring Activity Cliffs in Medicinal Chemistry Miniperspective.,* J. Med. Chem. 2012, 55, 2932–2942.

Stumpfe D., Hu Y., Dimova D., and Bajorath J., *Method for the Evaluation of Structure Activity Relationship Information Associated with Coordinated Activity Cliffs.,* J. Med. Chem., 2014, 57, 18-28.

Sutherland J. J., O'Brien L. A., Weaver D. F., *Spline-fitting with a genetic algorithm: A method for developing classification structure-activity relationships.,* J. Chem. Inf. Comp. Sci., 2003, 43, 1906-1915.

Tanimoto T. T., IBM Internal Report, 17th Nov 1957.

Taylor R. D., MacCoss M., and Lawson A. D. G., *Rings in Drugs*., J. Med. Chem., 2014, 57, 5845-5859.

Topliss, J. G., *Manual method for applying hansch approach to drug design.,* J. Med. Chem., 1977, 20, 463–469.

Trapero A., Pasito A., Singh V., Sabbah M., Coyne A. G., Mizrahi V., Blundell T. L., Ascher D. B., and Abell C., *Fragment-Based Approach to Targeting Inosine-5′-monophosphate Dehydrogenase (IMPDH) from Mycobacterium tuberculosis.,* J. Med. Chem., 2018, 61, 2806-2822.

Tropsha A., *Best Practices for QSAR Model Development, Validation, and Exploitation.,* Mol., Inf., 2010, 29, 476-488.

Trott O., and Olson A. J., *Software News and Update AutoDock Vina: Improving the Speed and Accuracy of Docking with a New Scoring Function, Efficient Optimization, and Multithreading.,* J. Comp. Chem., 2010, 31, 455-461.

Truchon J-F. and Bayly C. I., *Evaluating virtual screening methods: Good and bad metrics for the "early recognition" problem.,* J. Chem. Inf. Model., 2007, 47, 488-508.

van Deursen R., Blum L. C., Lorenz C., Reymond J. L., *A Searchable Map of PubChem.,* J. Chem. Inf. Model., 2010, 50, 1924-1934.

van Rossum G., Python tutorial, Technical Report CS-R9526, Centrum voor Wiskunde en Informatica (CWI), Amsterdam, May 1995.; Python Software Foundation. Python Language Reference, version 2.7. Available at http://www.python.org. Accessed on 1 September 2015.

Varin T., Bureau R., Mueller C. and Willett P., *Clustering files of chemical structures using the Szekely-Rizzo generalization of Ward's method.,* J. Mol. Graph. Model., 2009, 28, 187-195.

Venere M., Horbinski C., Crish J. F., Jin X., Vasanji A., Major J., Burrows A. C., Chang C., Prokop J., Wu Q. L., Sims P. A., Canoll P., Summers M. K., Rosenfeld S. S., and Rich J. N., *The mitotic kinesin KIF11 is a driver of invasion, proliferation, and self-renewal in glioblastoma.,* Sci. Trans. Med., 2015, 7, 304ra143.

Wade R., Clark K. J. and Goodford P. J., *Clustering files of chemical structures using the Szekely-Rizzo generalization of Ward's method.,* J. Med. Chem., 1993, 36, 140-147.

Wade R., and Goodford P. J., *Further development of hydrogen bond functions for use in determining energetically favorable binding sites on molecules of known structure. 2. Ligand probe groups with the ability to form more than two hydrogen bonds.,* J. Med. Chem., 1993, 36, 148-156.

Wang G., and Zhu W., *Molecular docking for drug discovery and development: a widely used approach but far from perfect.,* Future Med. Chem., 2016, 8, 1707-1710.

Wang F., Liu H., Blanton W. P., Belkina A., Lebrasseur N. K., and Denis G. V., *Brd2 disruption in mice causes severe obesity without Type 2 diabetes.,* The Biochemical Journal, 2010, 425, 71–83.

Weaver S., and Gleeson M. P., *The importance of the domain of applicability in QSAR modelling*, J. Mol. Graphics Model., 2008, 26, 1315-1326.

Wetzel S., Klein K., Renner S., Rauh D., Oprea T. I., Mutzel P., and Waldmann H., *Interactive exploration of chemical space with Scaffold Hunter.,* Nature Chem. Bio., 2009, 5, 581-583.
Willett P., *Dissimilarity-based algorithms for selecting structurally diverse sets of compounds.,* J. Comput. Biol, 1999, 6, 447-457.

Willett P., *The Calculation of Molecular Structural Similarity: Principles and Practice.,* Mol. Inf., 2014, 33, 403-413.

Willett P., J., Bardnard J. M., Downs G. M., *Chemical Similarity Searching.,* Chem. Inf. Comput. Sci., 1998, 38, 983-996.

Winter G., and McAuley K. E., *Automated data collection for macromolecular crystallography*, Methods, 2011, 55, 81-93.

Xu Y. J. and Johnson M., *Algorithm for Naming Molecular Equivalence Classes Represented by Labeled Pseudographs.,* J. Chem. Inf. Comput. Sci., 2001, 41, 181-185.

Yabucchi H., enrichvs v0.0.5, 2011, Available at https://CRAN.R-project.org/package=enrichvs

Yan A., Grant G. H., and Richards W. G., *Dynamics of conserved waters in human Hsp90: implications for drug design.,* J. R. Soc. Interface., 2008, 5, S199-S205.

Yang Y., Xu Z. J., Zhang Z. Y., Yang Z., Liu Y. T., Wang J. N., Cai T. T., Li S. J., Chen K. X., Shi J. Y., and Zhu W. L., *Like-Charge Guanidinium Pairing between Ligand and Receptor: An Unusual Interaction for Drug Discovery and Design?,* 2015, J. Phys. Chem. B., 119, 11988-11997.

Zhou H., and Skolnick J., J. Chem. Inf. Model., *FINDSITEcomb: A Threading/Structure-Based, Proteomic-Scale Virtual Ligand Screening Approach.,* 2013, 53, 230-240.

# Appendix A

# Figures

## Chapter 3

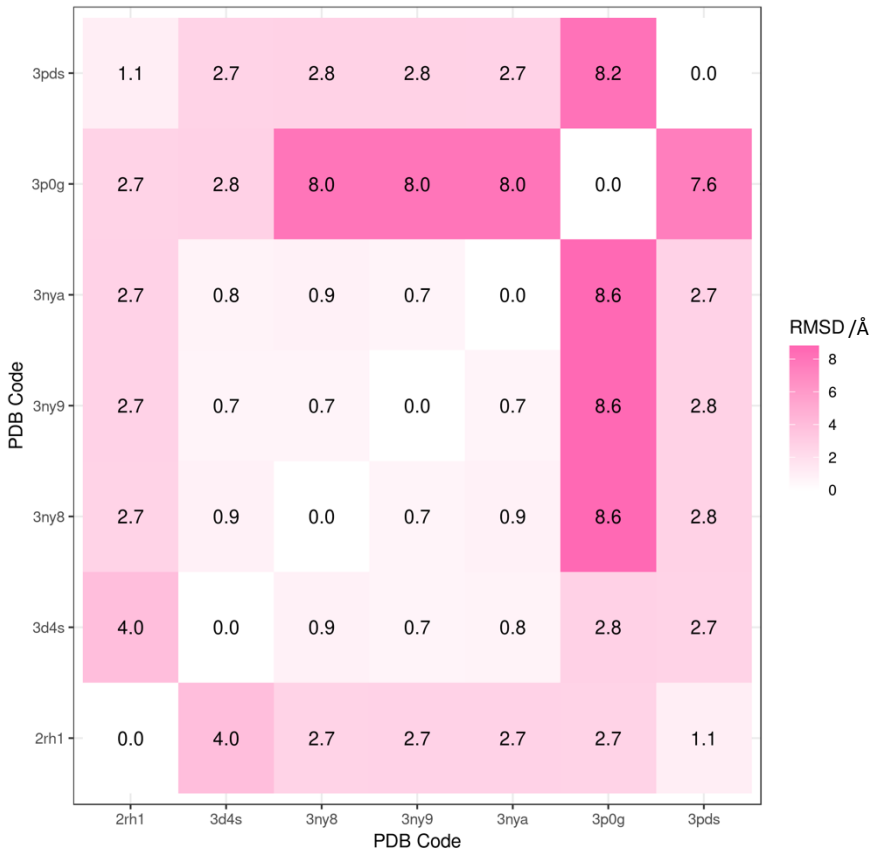



Figure A.3.1. All-atom RMSD of the PDB structures for the ADRB2 DUD-E dataset.

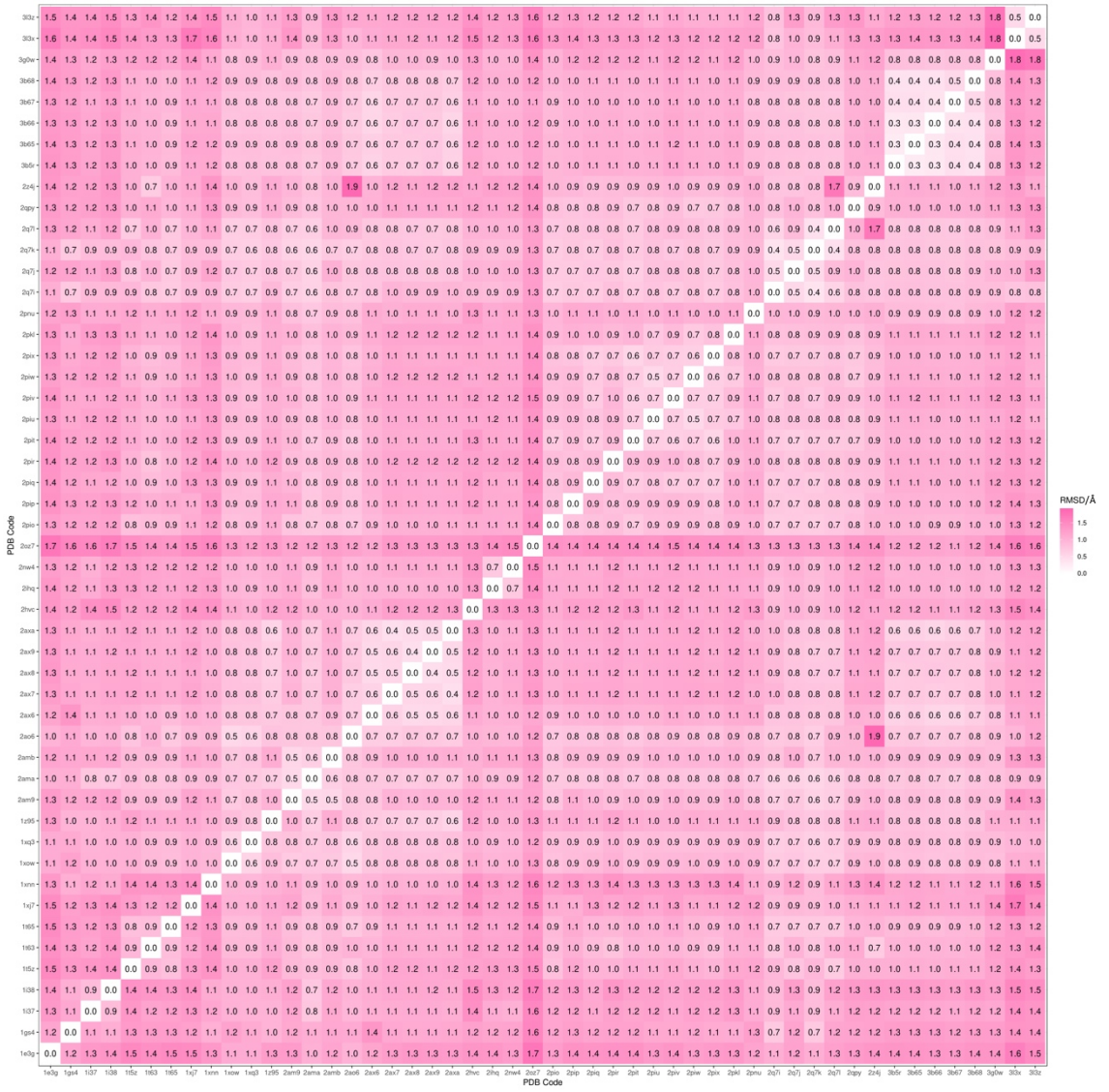



Figure A.3.2. All-atom RMSD of the PDB structures for the ANDR DUD-E dataset.

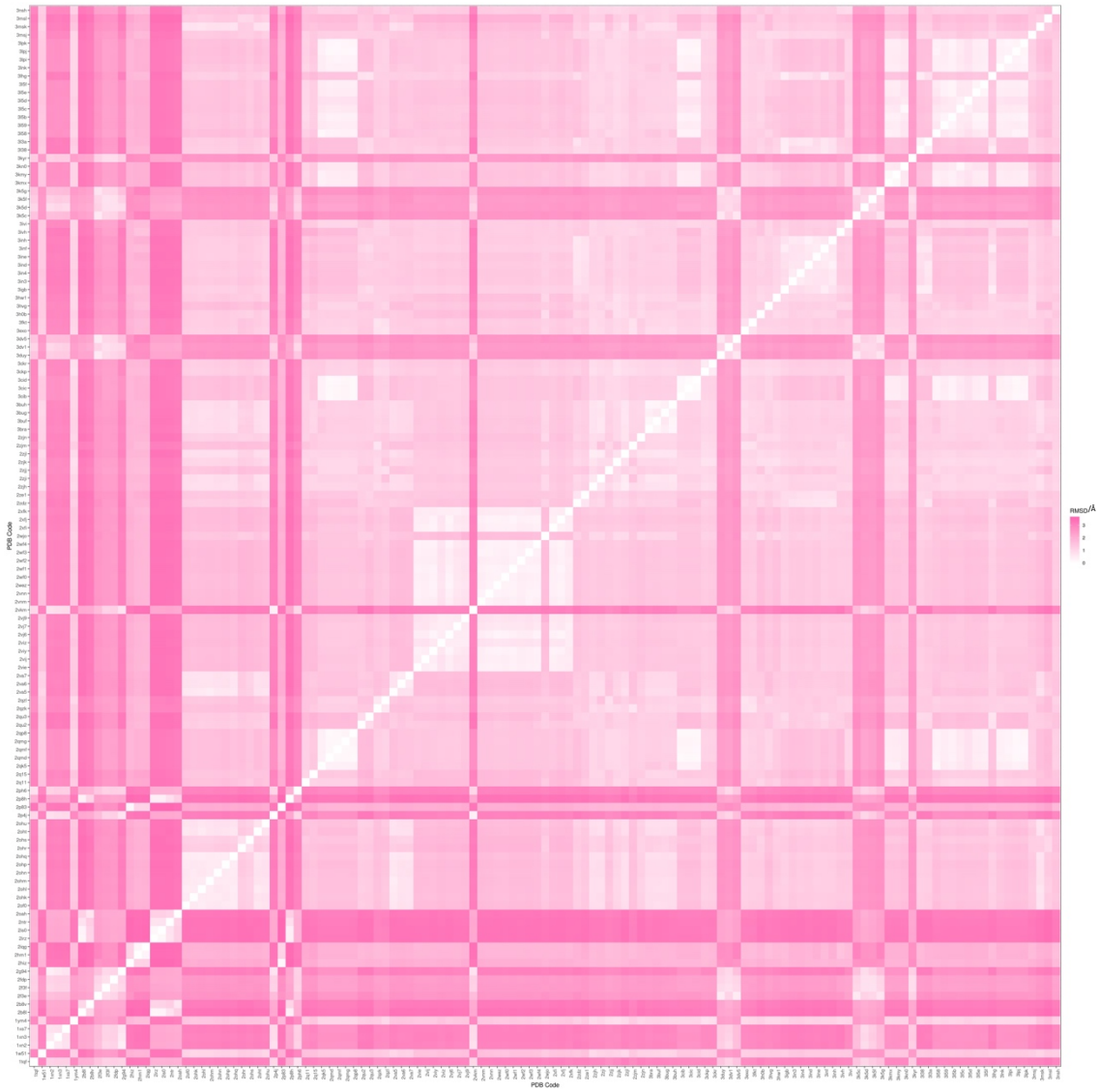



Figure A.3.3. All-atom RMSD of the PDB structures for the BACE1 DUD-E dataset.

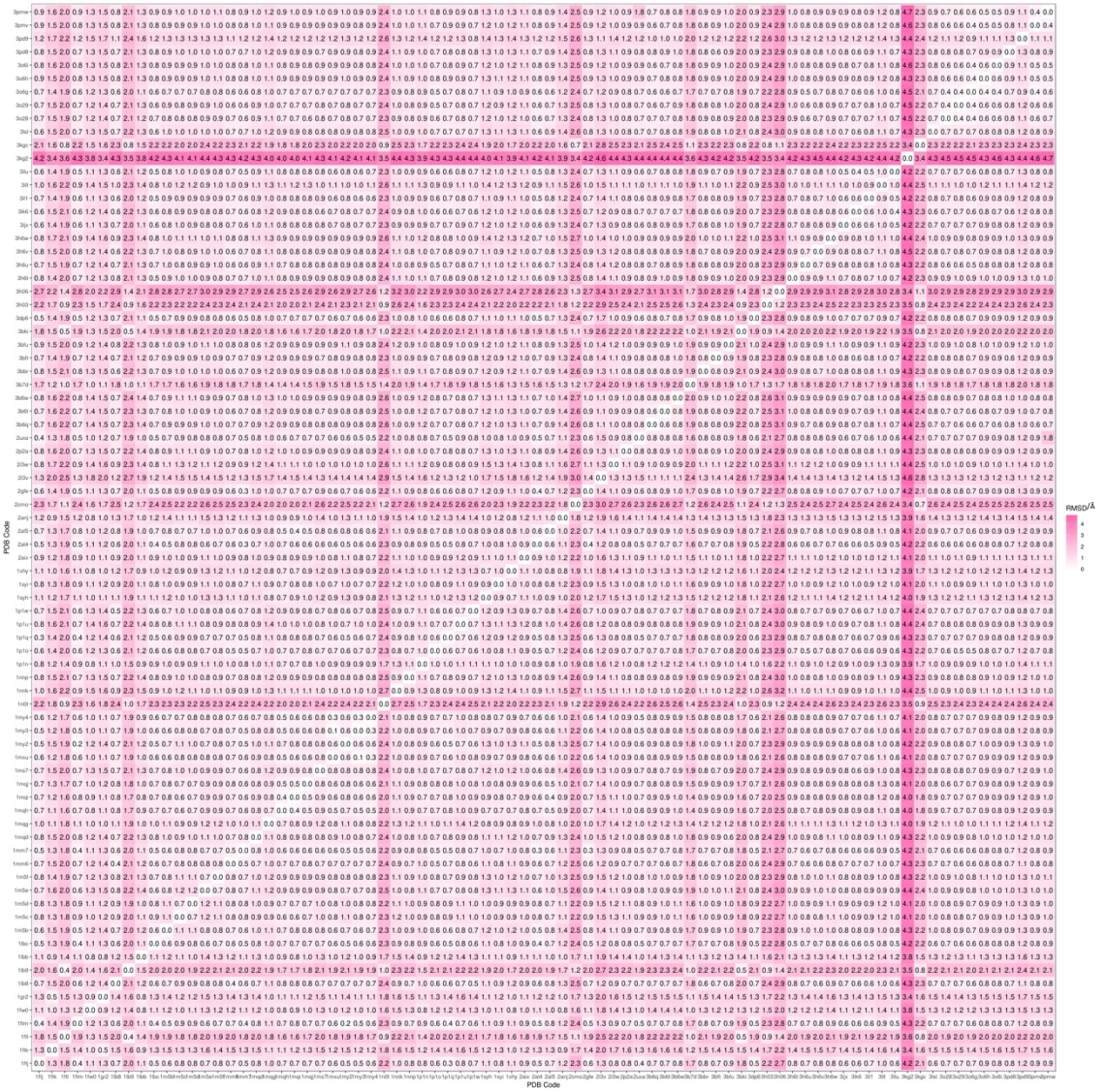



Figure A.3.4. All-atom RMSD of the PDB structures for the GRIA2 DUD-E dataset.

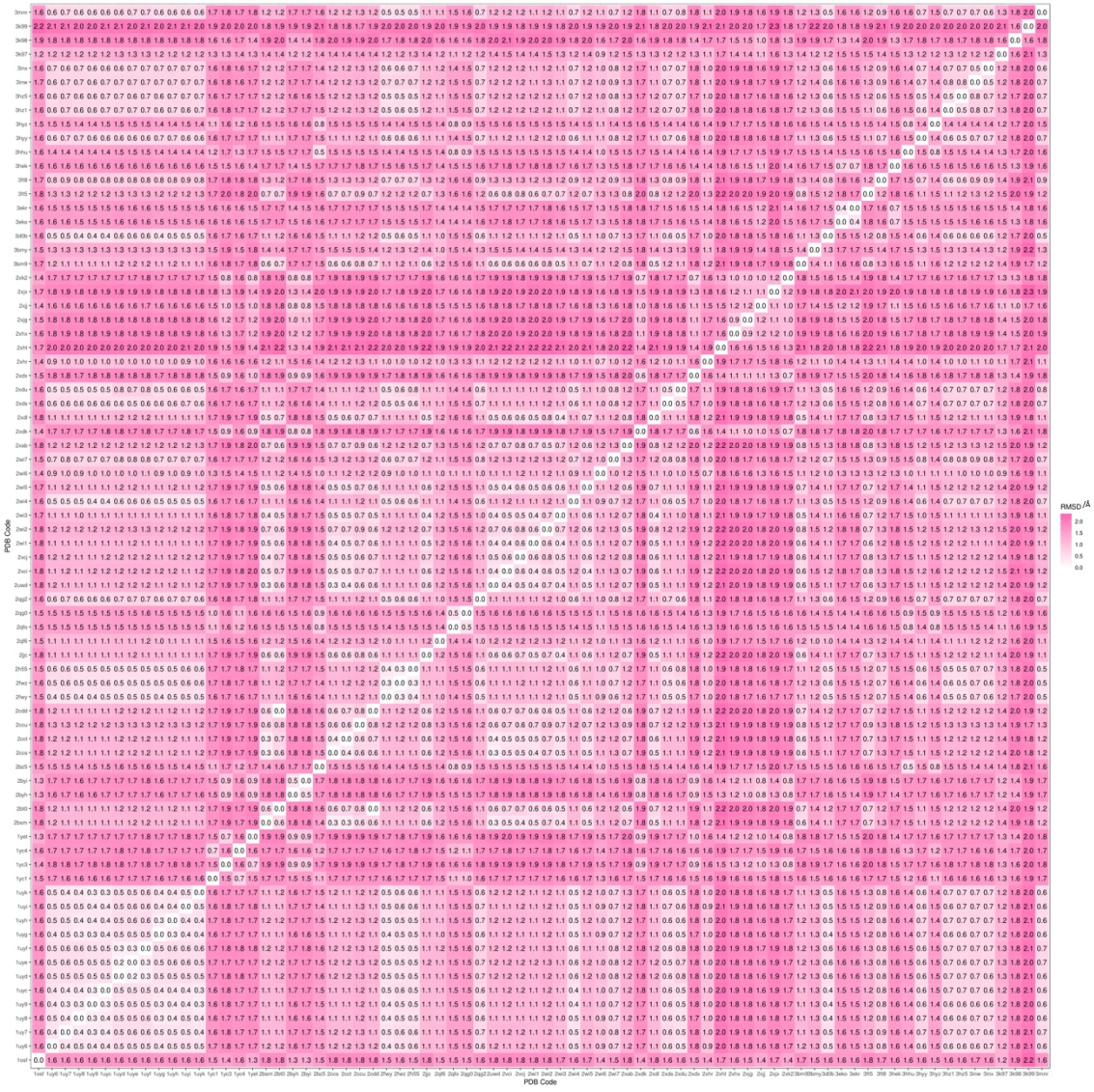



Figure A.3.5. All-atom RMSD of the PDB structures for the HSP90A DUD-E dataset.

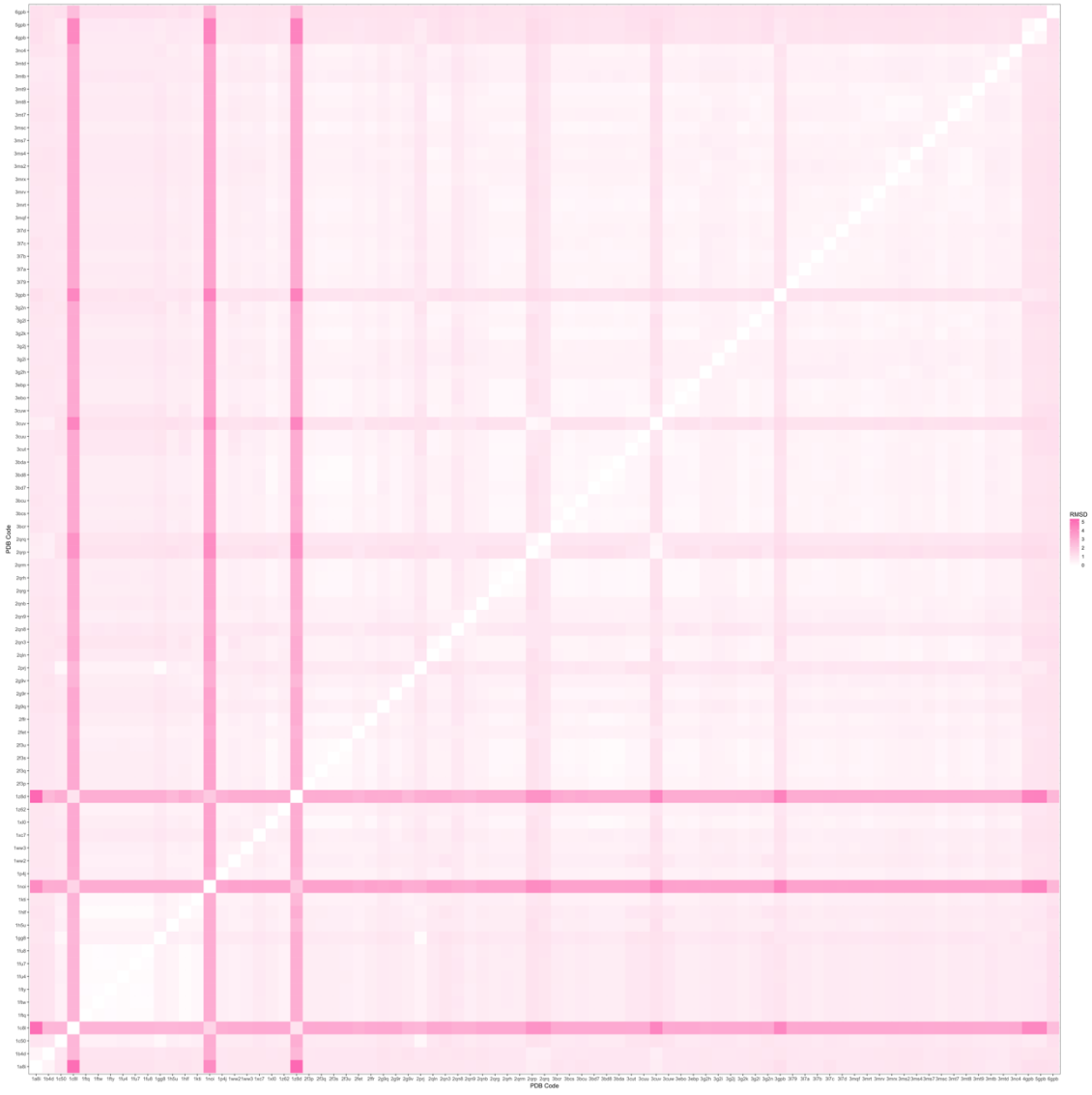



Figure A.3.6. All-atom RMSD of the PDB structures for the PYGM DUD-E dataset.

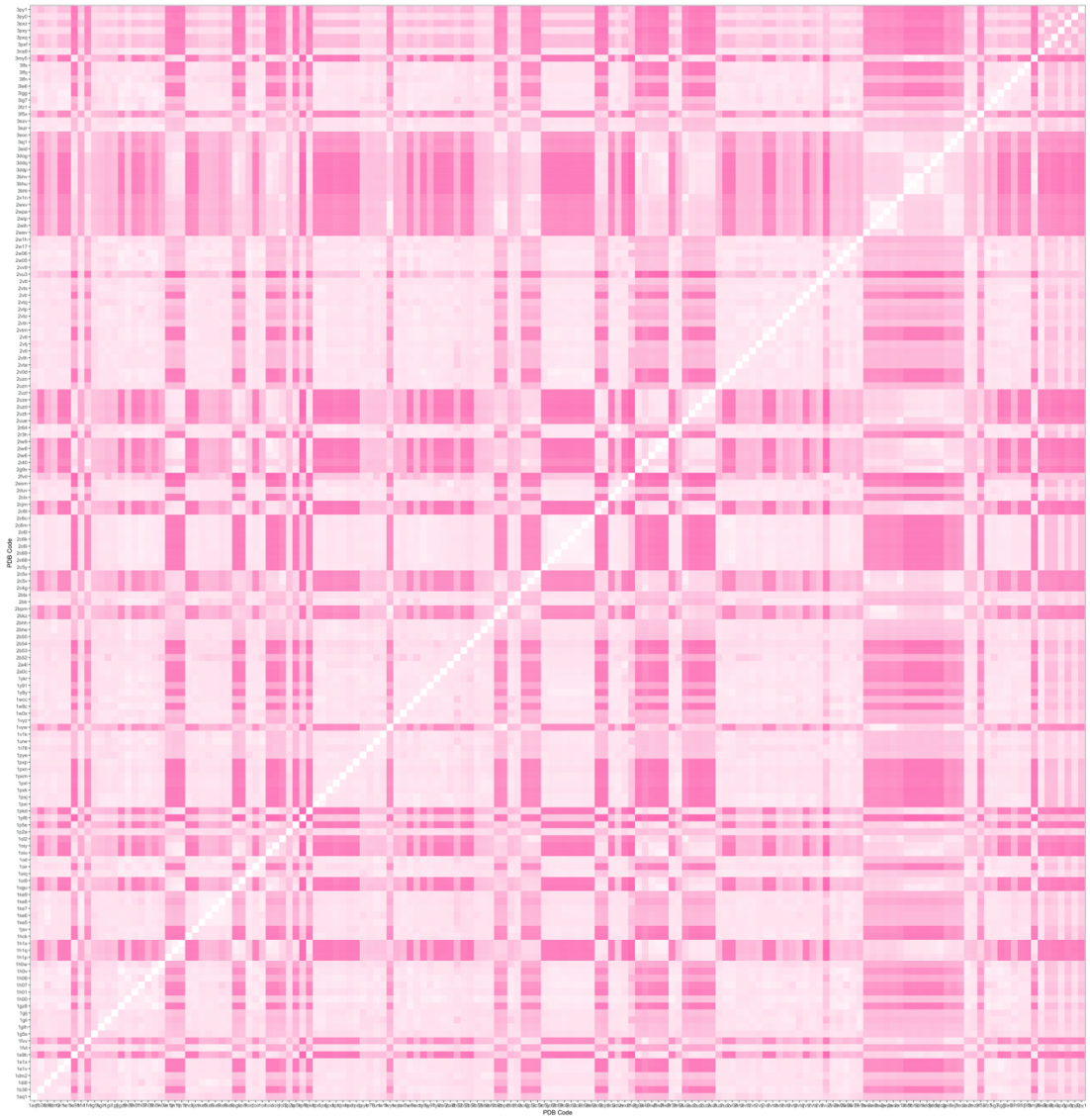



Figure A.3.7. All-atom RMSD of the PDB structures for the CDK2 dataset.

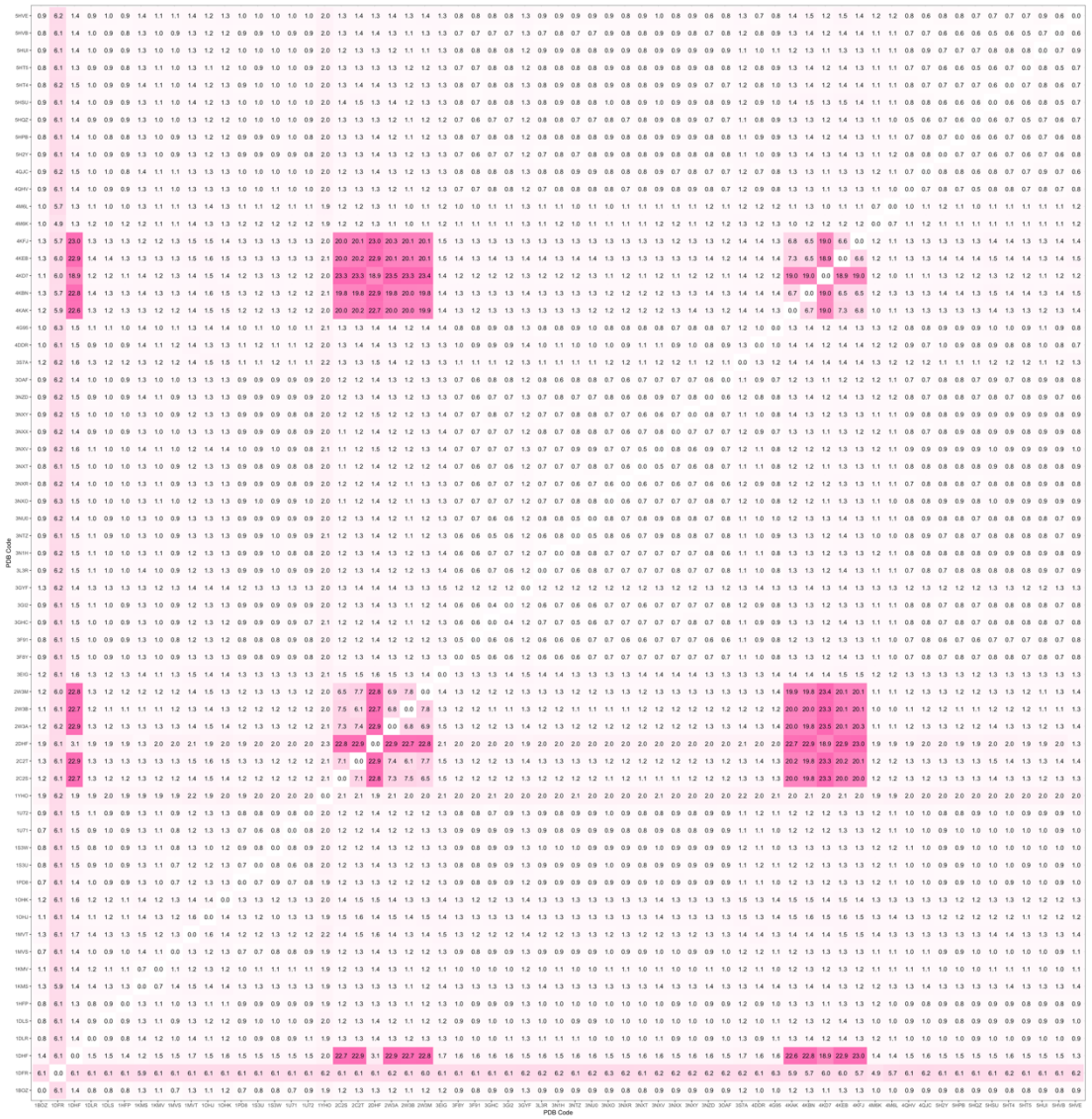



Figure A.3.8. All-atom RMSD of the PDB structures for the DHFR dataset.

| | | ADRB2 | | ANDR | | BACE1 | | GRIA2 | | HSP90A | | PYGM | | BRD1 | | CDK2 | | DHFR | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Max. | Min. | Max. | Min. | Max. | Min. | Max. | Min. | Max. | Min. | Max. | Min. | Max. | Min. | Max. | Min. | Max. | Min. |
| 5 structures | Morgan Fingerprint Score | 0.91 | 0.80 | 0.91 | 0.72 | 0.82 | 0.69 | 0.68 | 0.59 | 0.98 | 0.61 | 0.96 | 0.81 | 0.60 | 0.41 | 0.54 | 0.49 | 0.81 | 0.77 |
| | 3D Pharmacophore Score | 0.75 | 0.59 | **0.85** | 0.58 | **0.81** | **0.66** | **0.77** | **0.55** | 0.87 | **0.81** | 0.60 | **0.56** | **0.61** | **0.52** | 0.60 | **0.54** | **0.87** | **0.69** |
| | CRANkS Interaction Score | 0.86 | 0.62 | 0.70 | 0.58 | 0.76 | 0.54 | 0.64 | 0.40 | **0.92** | 0.56 | **0.69** | 0.43 | 0.51 | 0.44 | **0.62** | 0.51 | 0.59 | 0.45 |
| | CRANkS Element Score | 0.87 | **0.65** | 0.81 | **0.60** | 0.79 | 0.52 | 0.76 | 0.41 | 0.87 | 0.60 | 0.68 | 0.36 | 0.59 | 0.33 | 0.59 | 0.47 | 0.70 | 0.54 |
| | CRANkS Pharmacophore Score | **0.89** | **0.65** | 0.76 | 0.56 | 0.79 | 0.47 | 0.61 | 0.48 | 0.84 | 0.60 | 0.63 | 0.43 | 0.60 | 0.37 | 0.57 | 0.49 | 0.78 | 0.58 |
| 10 structures | Morgan Fingerprint Score | N/A | N/A | 0.93 | 0.72 | 0.84 | 0.78 | 0.75 | 0.67 | 0.89 | 0.54 | 0.96 | 0.88 | 0.56 | 0.40 | 0.60 | 0.50 | 0.84 | 0.75 |
| | 3D Pharmacophore Score | N/A | N/A | **0.81** | **0.63** | 0.85 | 0.76 | **0.72** | **0.68** | 0.82 | 0.73 | 0.59 | **0.55** | **0.59** | **0.52** | **0.64** | **0.53** | **0.85** | **0.69** |
| | CRANkS Interaction Score | N/A | N/A | 0.74 | 0.61 | 0.77 | 0.60 | 0.60 | 0.35 | **0.91** | **0.80** | 0.64 | 0.45 | 0.51 | 0.46 | 0.60 | 0.49 | 0.75 | 0.34 |
| | CRANkS Element Score | N/A | N/A | 0.73 | 0.53 | 0.86 | 0.76 | **0.72** | 0.47 | 0.81 | 0.53 | **0.69** | 0.48 | 0.50 | 0.35 | 0.59 | 0.46 | 0.65 | 0.44 |
| | CRANkS Pharmacophore Score | N/A | N/A | 0.69 | 0.52 | **0.87** | **0.77** | 0.67 | 0.55 | 0.82 | 0.45 | 0.61 | 0.40 | 0.52 | 0.36 | 0.54 | 0.41 | 0.73 | 0.60 |
| 25 structures | Morgan Fingerprint Score | N/A | N/A | N/A | N/A | 0.87 | 0.82 | 0.85 | 0.75 | 0.95 | 0.65 | 0.93 | 0.87 | 0.46 | 0.38 | 0.57 | 0.50 | 0.84 | 0.80 |
| | 3D Pharmacophore Score | N/A | N/A | N/A | N/A | 0.86 | 0.80 | 0.81 | **0.66** | 0.85 | 0.73 | 0.59 | 0.56 | 0.55 | **0.50** | **0.61** | **0.58** | **0.86** | **0.80** |
| | CRANkS Interaction Score | N/A | N/A | N/A | N/A | 0.87 | 0.73 | 0.62 | 0.53 | 0.89 | 0.68 | 0.66 | 0.37 | 0.56 | 0.43 | 0.58 | 0.49 | 0.69 | 0.62 |
| | CRANkS Element Score | N/A | N/A | N/A | N/A | 0.87 | 0.81 | 0.58 | 0.52 | 0.91 | 0.75 | **0.68** | **0.57** | 0.53 | 0.38 | 0.58 | 0.52 | 0.66 | 0.55 |
| | CRANkS Pharmacophore Score | N/A | N/A | N/A | N/A | **0.90** | **0.83** | 0.59 | 0.55 | **0.93** | **0.79** | 0.58 | 0.43 | **0.57** | 0.40 | 0.56 | 0.44 | 0.74 | 0.68 |
| 50 structures | Morgan Fingerprint Score | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | 0.93 | 0.86 | 0.94 | 0.93 | N/A | N/A | 0.57 | 0.51 | **0.84** | 0.73 |
| | 3D Pharmacophore Score | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | 0.82 | 0.81 | 0.59 | 0.57 | N/A | N/A | **0.60** | **0.55** | 0.81 | **0.73** |
| | CRANkS Interaction Score | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | 0.94 | 0.80 | 0.64 | 0.40 | N/A | N/A | 0.58 | 0.53 | 0.70 | 0.60 |
| | CRANkS Element Score | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | 0.92 | 0.87 | **0.69** | **0.60** | N/A | N/A | 0.59 | 0.53 | 0.64 | 0.61 |
| | CRANkS Pharmacophore Score | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | **0.95** | **0.88** | 0.56 | 0.46 | N/A | N/A | 0.54 | 0.51 | 0.72 | 0.68 |

Figure A.3.9 Table to show the maximum and minimum AUC values for each target for different methods. The values are split by the number of input structures and are the maximum and minimum across the five separate runs of different input structures. The highest AUC values for each number of input structures is outlined by a black order and the highest value out of the 3D Pharmacophore Fingerprint Score and CRANkS scores is shown in bold.

| | | ADRB2 | | ANDR | | BACE1 | | GRIA2 | | HSP90A | | PYGM | | BRD1 | | CDK2 | | DHFR | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Max. | Min. | Max. | Min. | Max. | Min. | Max. | Min. | Max. | Min. | Max. | Min. | Max. | Min. | Max. | Min. | Max. | Min. |
| 5 structures | Morgan Fingerprint Score | 0.65 | 0.46 | 0.74 | 0.51 | 0.70 | 0.07 | 0.32 | 0.20 | 0.88 | 0.28 | 0.52 | 0.14 | 0.19 | 0.04 | 0.09 | 0.01 | 0.81 | 0.13 |
| | 3D Pharmacophore Score | 0.51 | 0.24 | **0.54** | **0.13** | **0.65** | **0.08** | 0.25 | **0.12** | **0.78** | **0.54** | 0.14 | 0.08 | 0.27 | 0.14 | 0.11 | 0.01 | **0.82** | **0.26** |
| | CRANkS Interaction Score | 0.35 | 0.15 | 0.17 | 0.08 | 0.42 | 0.01 | 0.16 | 0.05 | 0.72 | 0.05 | 0.03 | 0.00 | **0.49** | **0.15** | 0.11 | 0.01 | 0.22 | 0.00 |
| | CRANkS Element Score | **0.47** | **0.27** | 0.31 | 0.12 | 0.64 | 0.01 | **0.39** | 0.06 | 0.51 | 0.18 | **0.16** | **0.10** | 0.17 | 0.02 | **0.20** | 0.03 | 0.34 | 0.02 |
| | CRANkS Pharmacophore Score | 0.46 | 0.23 | 0.33 | **0.13** | 0.63 | 0.02 | 0.31 | 0.05 | 0.40 | 0.04 | 0.11 | 0.03 | 0.19 | 0.07 | 0.14 | **0.04** | 0.61 | 0.24 |
| 10 structures | Morgan Fingerprint Score | N/A | N/A | 0.74 | 0.45 | 0.73 | 0.57 | 0.47 | 0.32 | 0.85 | 0.33 | 0.49 | 0.16 | 0.21 | 0.01 | 0.08 | 0.03 | 0.72 | 0.34 |
| | 3D Pharmacophore Score | N/A | N/A | **0.50** | **0.17** | **0.68** | 0.53 | 0.23 | 0.16 | 0.64 | **0.44** | 0.13 | 0.08 | 0.31 | **0.12** | 0.11 | 0.01 | **0.82** | **0.31** |
| | CRANkS Interaction Score | N/A | N/A | 0.22 | 0.09 | 0.22 | 0.01 | 0.09 | 0.07 | **0.75** | 0.16 | 0.01 | 0.00 | **0.46** | 0.10 | 0.05 | 0.03 | 0.25 | 0.00 |
| | CRANkS Element Score | N/A | N/A | 0.20 | 0.05 | 0.67 | **0.55** | **0.40** | 0.11 | 0.61 | 0.09 | **0.14** | **0.10** | 0.10 | 0.01 | 0.10 | **0.04** | 0.11 | 0.00 |
| | CRANkS Pharmacophore Score | N/A | N/A | 0.23 | 0.11 | 0.67 | 0.41 | 0.29 | **0.18** | 0.47 | 0.18 | 0.10 | 0.05 | 0.12 | 0.04 | **0.13** | 0.03 | 0.33 | 0.18 |
| 25 structures | Morgan Fingerprint Score | N/A | N/A | N/A | N/A | 0.74 | 0.70 | 0.51 | 0.43 | 0.83 | 0.57 | 0.35 | 0.21 | 0.16 | 0.06 | 0.15 | 0.09 | 0.77 | 0.42 |
| | 3D Pharmacophore Score | N/A | N/A | N/A | N/A | 0.69 | 0.60 | **0.22** | **0.16** | 0.71 | 0.41 | 0.14 | 0.09 | 0.25 | **0.15** | 0.07 | 0.03 | **0.57** | **0.25** |
| | CRANkS Interaction Score | N/A | N/A | N/A | N/A | 0.39 | 0.17 | 0.16 | 0.11 | 0.51 | 0.30 | 0.02 | 0.00 | **0.50** | 0.10 | 0.13 | 0.02 | 0.35 | 0.03 |
| | CRANkS Element Score | N/A | N/A | N/A | N/A | 0.68 | **0.62** | 0.20 | 0.10 | 0.71 | 0.40 | 0.15 | 0.10 | 0.14 | 0.07 | 0.20 | 0.05 | 0.13 | 0.01 |
| | CRANkS Pharmacophore Score | N/A | N/A | N/A | N/A | **0.71** | 0.61 | 0.20 | 0.11 | **0.87** | **0.62** | **0.16** | **0.12** | 0.11 | 0.04 | **0.23** | **0.09** | 0.35 | 0.06 |
| 50 structures | Morgan Fingerprint Score | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | 0.87 | 0.73 | 0.35 | 0.27 | N/A | N/A | 0.13 | 0.11 | 0.58 | 0.27 |
| | 3D Pharmacophore Score | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | 0.65 | 0.51 | 0.13 | **0.12** | N/A | N/A | 0.08 | 0.03 | 0.38 | 0.27 |
| | CRANkS Interaction Score | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | 0.60 | 0.49 | 0.01 | 0.01 | N/A | N/A | 0.11 | 0.02 | 0.35 | 0.24 |
| | CRANkS Element Score | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | 0.67 | 0.41 | **0.15** | **0.12** | N/A | N/A | 0.19 | 0.02 | 0.03 | 0.00 |
| | CRANkS Pharmacophore Score | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | **0.85** | **0.70** | **0.15** | **0.12** | N/A | N/A | **0.30** | **0.05** | **0.40** | **0.28** |

Figure A.3.10 Table to show the maximum and minimum BEDROC values for each target for different methods. The values are split by the number of input structures and are the maximum and minimum across the five separate runs of different input structures. The highest BEDROC values for each number of input structures is outlined by a black order and the highest value out of the 3D Pharmacophore Fingerprint Score and CRANkS scores is shown in bold. Notably for target BRD1 the CRANkS Interaction Score achieves much higher maximum scores than the fingerprint scores and for target CDK2 the CRANkS Pharmacophore Score shows the same behaviour. Again, the CRANkS scores exhibit much larger ranges than the fingerprint scores.

| | | ADRB2 | | ANDR | | BACE1 | | GRIA2 | | HSP90A | | PYGM | | BRD1 | | CDK2 | | DHFR | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Max. | Min. | Max. | Min. | Max. | Min. | Max. | Min. | Max. | Min. | Max. | Min. | Max. | Min. | Max. | Min. | Max. | Min. |
| 5 structures | Morgan Fingerprint Score | 46 | 29 | 48 | 29 | 55 | 4 | 22 | 12 | 60 | 13 | 29 | 8 | 1 | 0 | 2 | 0 | 15 | 1 |
| | 3D Pharmacophore Score | **35** | 15 | **34** | **7** | 47 | **8** | 18 | **7** | **58** | **33** | 8 | 4 | **1** | 0 | 2 | 0 | **15** | **5** |
| | CRANkS Interaction Score | 20 | 10 | 9 | 4 | 28 | 0 | 13 | 4 | 39 | 1 | 1 | 0 | **1** | 0 | 3 | 0 | 1 | 0 |
| | CRANkS Element Score | 29 | **16** | 18 | 6 | **49** | 0 | **30** | 5 | 29 | 8 | **9** | **5** | 0 | 0 | **4** | **1** | 4 | 0 |
| | CRANkS Pharmacophore Score | 29 | 14 | 18 | 7 | 48 | 1 | 23 | 3 | 21 | 1 | 6 | 1 | 0 | 0 | 3 | **1** | 13 | 0 |
| 10 structures | Morgan Fingerprint Score | N/A | N/A | 47 | 25 | 57 | 38 | 34 | 20 | 53 | 18 | 26 | 8 | 1 | 0 | 2 | 1 | 15 | 5 |
| | 3D Pharmacophore Score | N/A | N/A | **30** | **9** | **52** | **36** | 13 | 10 | 35 | **27** | **8** | 4 | **1** | 0 | 2 | 0 | **15** | **6** |
| | CRANkS Interaction Score | N/A | N/A | 11 | 4 | 13 | 0 | 6 | 5 | **44** | 9 | 0 | 0 | **1** | 0 | 1 | 0 | 4 | 0 |
| | CRANkS Element Score | N/A | N/A | 10 | 2 | 49 | **36** | **29** | 6 | 33 | 4 | **8** | **5** | 0 | 0 | **3** | **1** | 2 | 0 |
| | CRANkS Pharmacophore Score | N/A | N/A | 13 | 6 | 51 | 26 | 19 | **12** | 28 | 11 | 5 | 4 | 0 | 0 | **3** | 0 | 4 | 0 |
| 25 structures | Morgan Fingerprint Score | N/A | N/A | N/A | N/A | 57 | 50 | 36 | 30 | 51 | 33 | 15 | 8 | 0 | 0 | 3 | 2 | 13 | 8 |
| | 3D Pharmacophore Score | N/A | N/A | N/A | N/A | **51** | 40 | **15** | **10** | 44 | 29 | 8 | 4 | **1** | 0 | 2 | 0 | **8** | **4** |
| | CRANkS Interaction Score | N/A | N/A | N/A | N/A | 20 | 9 | 13 | 8 | 33 | 14 | 0 | 0 | **1** | 0 | 2 | 0 | 2 | 0 |
| | CRANkS Element Score | N/A | N/A | N/A | N/A | 49 | **47** | 14 | 7 | 39 | 21 | 8 | **5** | 0 | 0 | 5 | 1 | 2 | 0 |
| | CRANkS Pharmacophore Score | N/A | N/A | N/A | N/A | 50 | 44 | 14 | 7 | **57** | **39** | **9** | **5** | 0 | 0 | **6** | **2** | 6 | 0 |
| 50 structures | Morgan Fingerprint Score | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | 63 | 47 | 15 | 11 | N/A | N/A | 4 | 1 | 12 | 4 |
| | 3D Pharmacophore Score | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | 49 | 28 | 7 | **6** | N/A | N/A | 2 | 0 | **8** | **6** |
| | CRANkS Interaction Score | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | 45 | 25 | 0 | 0 | N/A | N/A | 2 | 0 | 4 | 2 |
| | CRANkS Element Score | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | 44 | 29 | **8** | **6** | N/A | N/A | 2 | 0 | 0 | 0 |
| | CRANkS Pharmacophore Score | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | **66** | **36** | **8** | **6** | N/A | N/A | **7** | **1** | **8** | 3 |

Figure A.3.11 Table to show the maximum and minimum $EF_{1\%}$ values for each target for different methods. The values are split by the number of input structures and are the maximum and minimum across the five separate runs of different input structures. The highest $EF_{1\%}$ values for each number of input structures is outlined by a black order and the highest value out of the 3D Pharmacophore Fingerprint Score and CRANkS scores is shown in bold. The CRANkS score shows a much larger range in performance than the fingerprint methods. The results only improve with increasing number of input structures for two targets: HSP90A and BACE1.

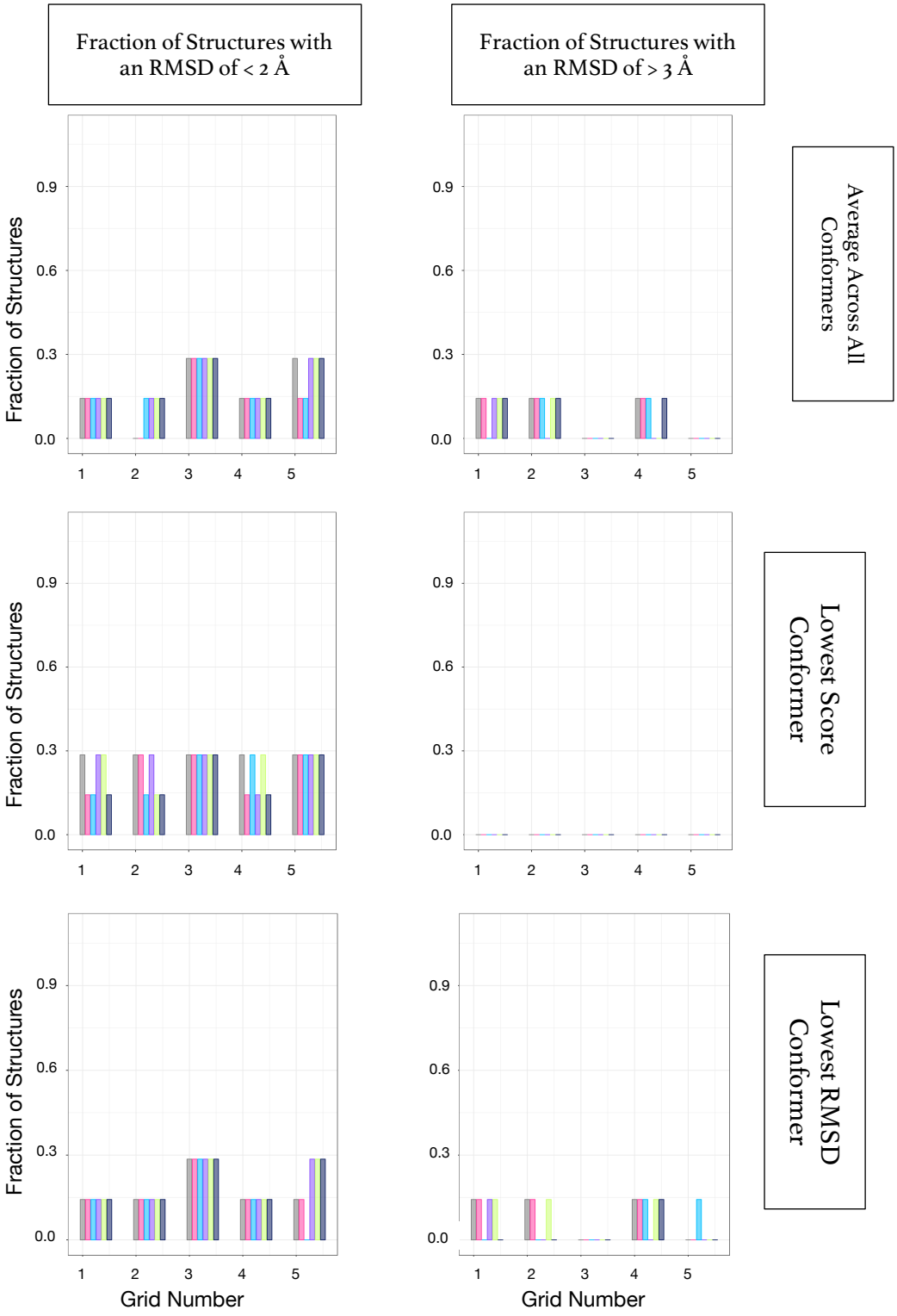



Figure A.4.1 Fraction of structures with a crystallographic RMSD of < 2 Å and an RMSD of > 3 Å are shown for the redocking of PDB structures for ADRB2. This is split into the average RMSD between the structure and all conformers, the RMSD between the structure and lowest scored conformer, and the lowest RMSD between the structure and ant conformer. The normalisation constant are coloured with grey for AutoDock, pink for 0.01, light blue for 0.1, purple for 1, green for 10 and dark blue for 100.

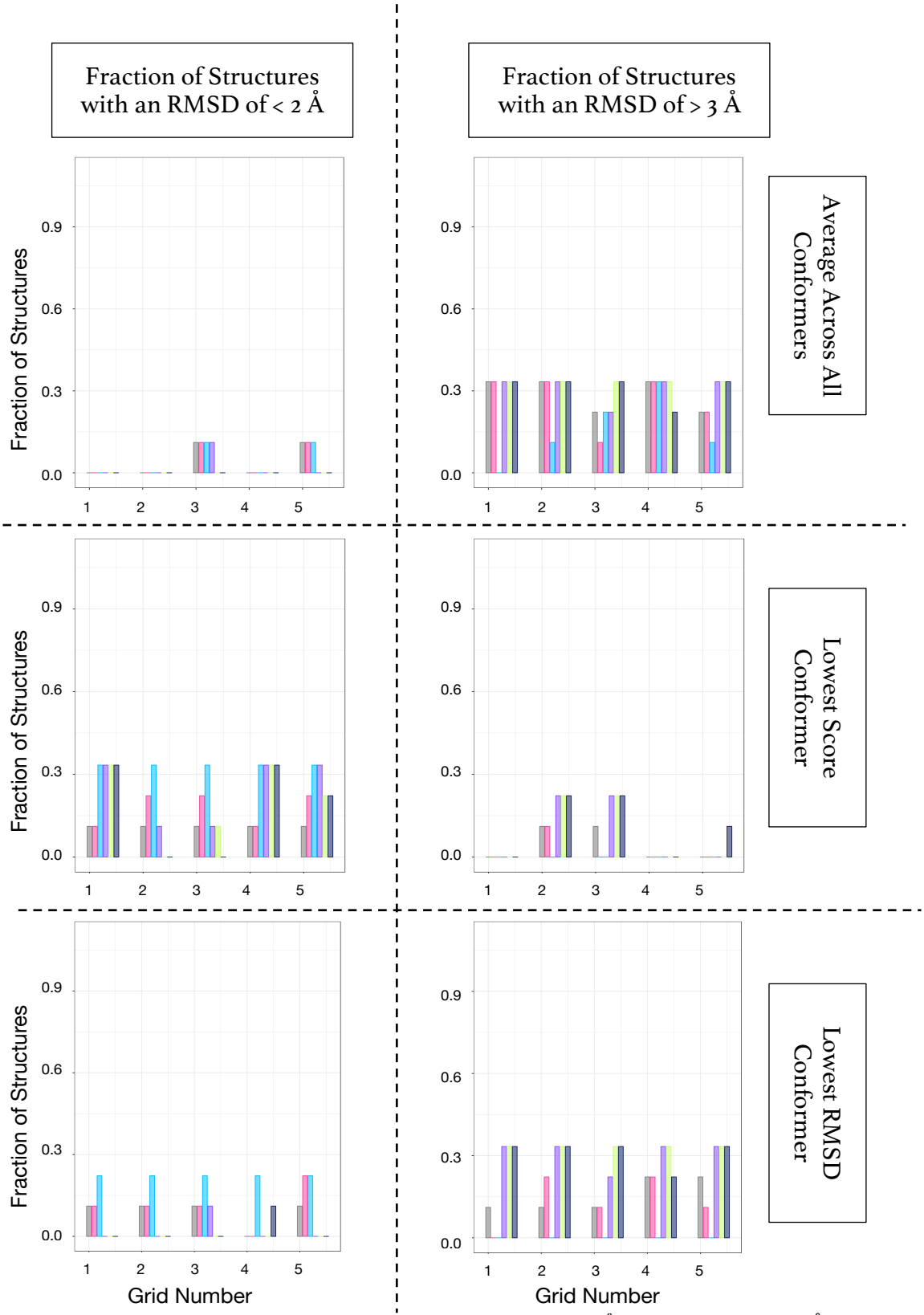



Figure A.4.2. Fraction of structures with a crystallographic RMSD of < 2 Å and an RMSD of > 3 Å are shown for the redocking of PDB structures for BCL2. This is split into the average RMSD between the structure and all conformers, the RMSD between the structure and lowest scored conformer, and the lowest RMSD between the structure and ant conformer. The various normalisation constants are coloured with grey for AutoDock, pink for 0.01, light blue for 0.1, purple for 1, green for 10 and dark blue for 100.

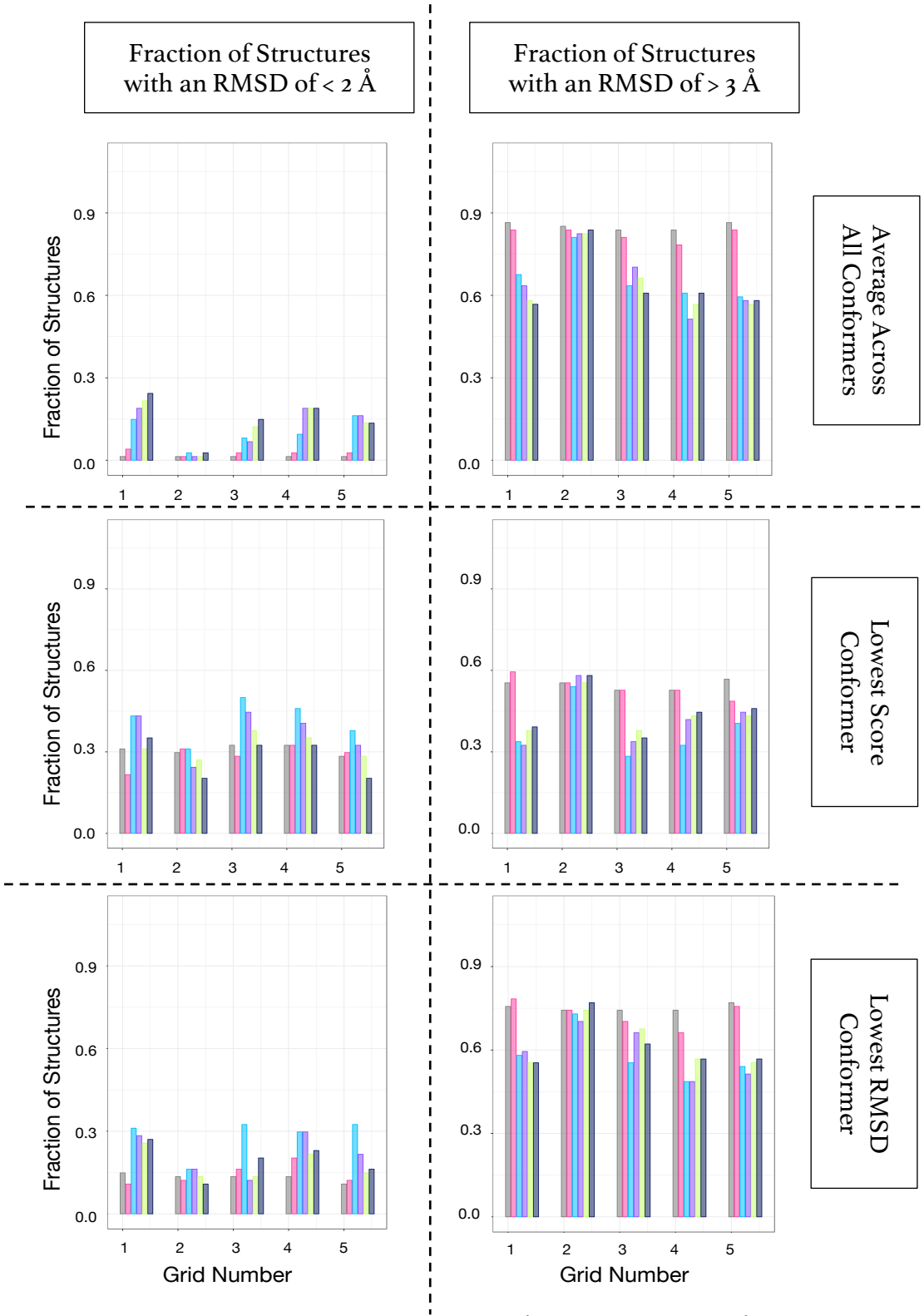



Figure A.4.3. Fraction of structures with an RMSD of < 2 Å and an RMSD of > 3 Å are shown for the redocking of PDB structures for HSP90A. This is split into the average RMSD between the structure and all conformers, the RMSD between the structure and lowest scored conformer, and the lowest RMSD between the structure and ant conformer. The normalisation constants are coloured grey for AutoDock, pink for AutoCRANkS $w$=0.01, light blue for $w$=0.1, purple for $w$=1, green for $w$=10 and dark blue for $w$=100.

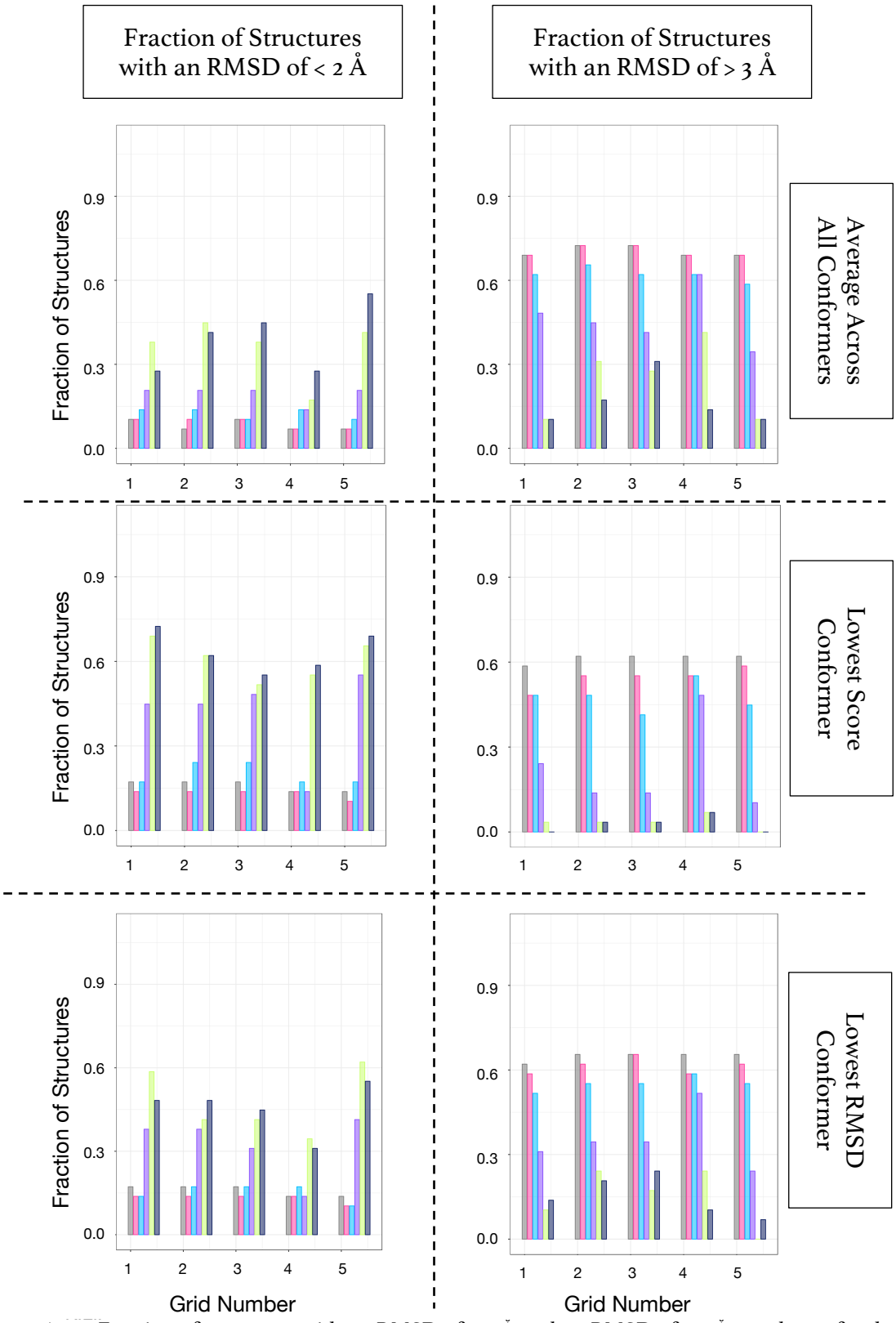



Figure A.4.4. Fraction of structures with an RMSD of < 2 Å and an RMSD of > 3 Å are shown for the redocking of PDB structures for KIF11. This is split into the average RMSD between the structure and all conformers, the RMSD between the structure and lowest scored conformer, and the lowest RMSD between the structure and ant conformer. The normalisation constants are coloured with grey for AutoDock, pink for 0.01, light blue for 0.1, purple for 1, green for 10 and dark blue for 100.

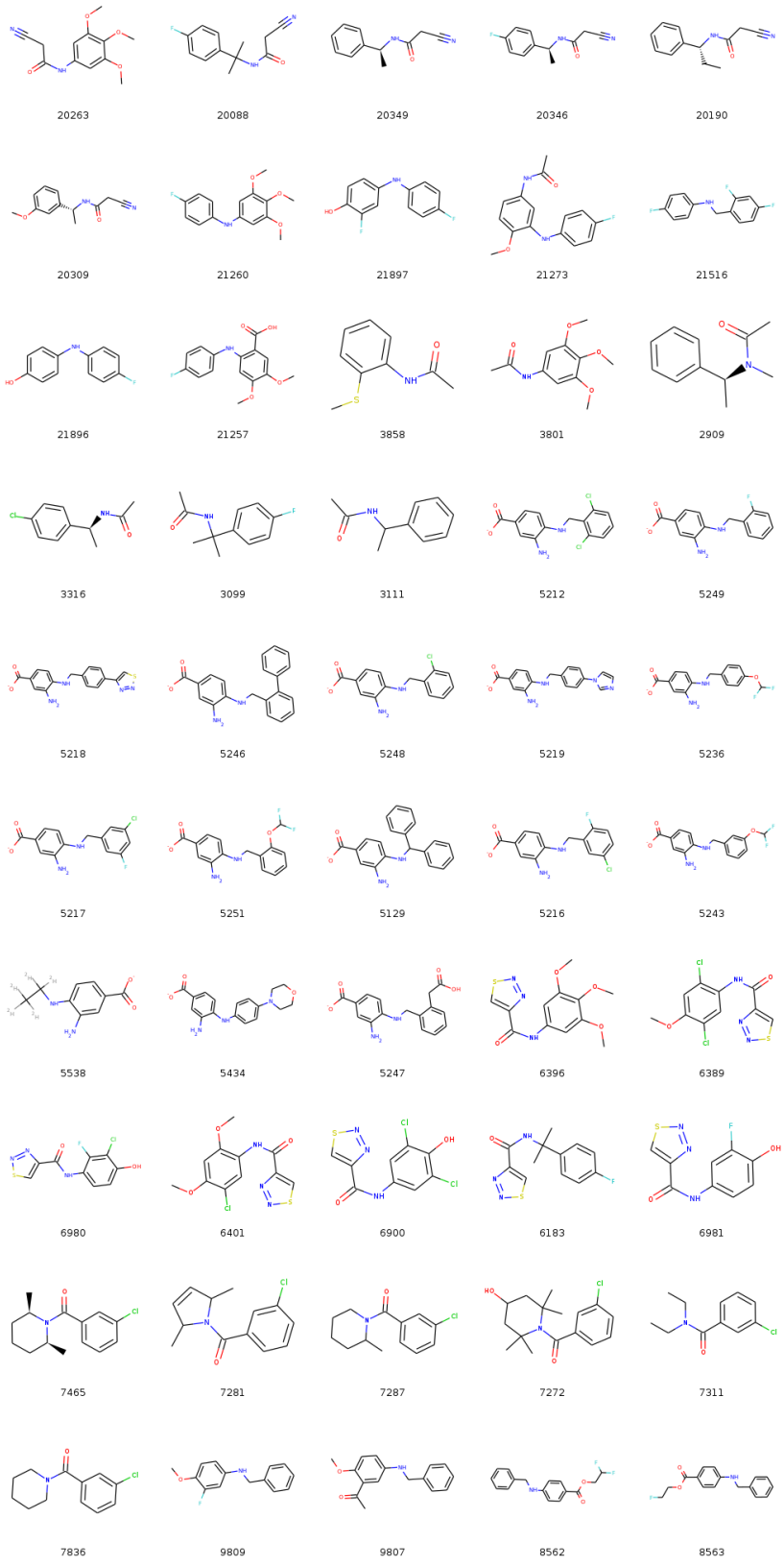



Figure A.5.1 (*continued on next page*)

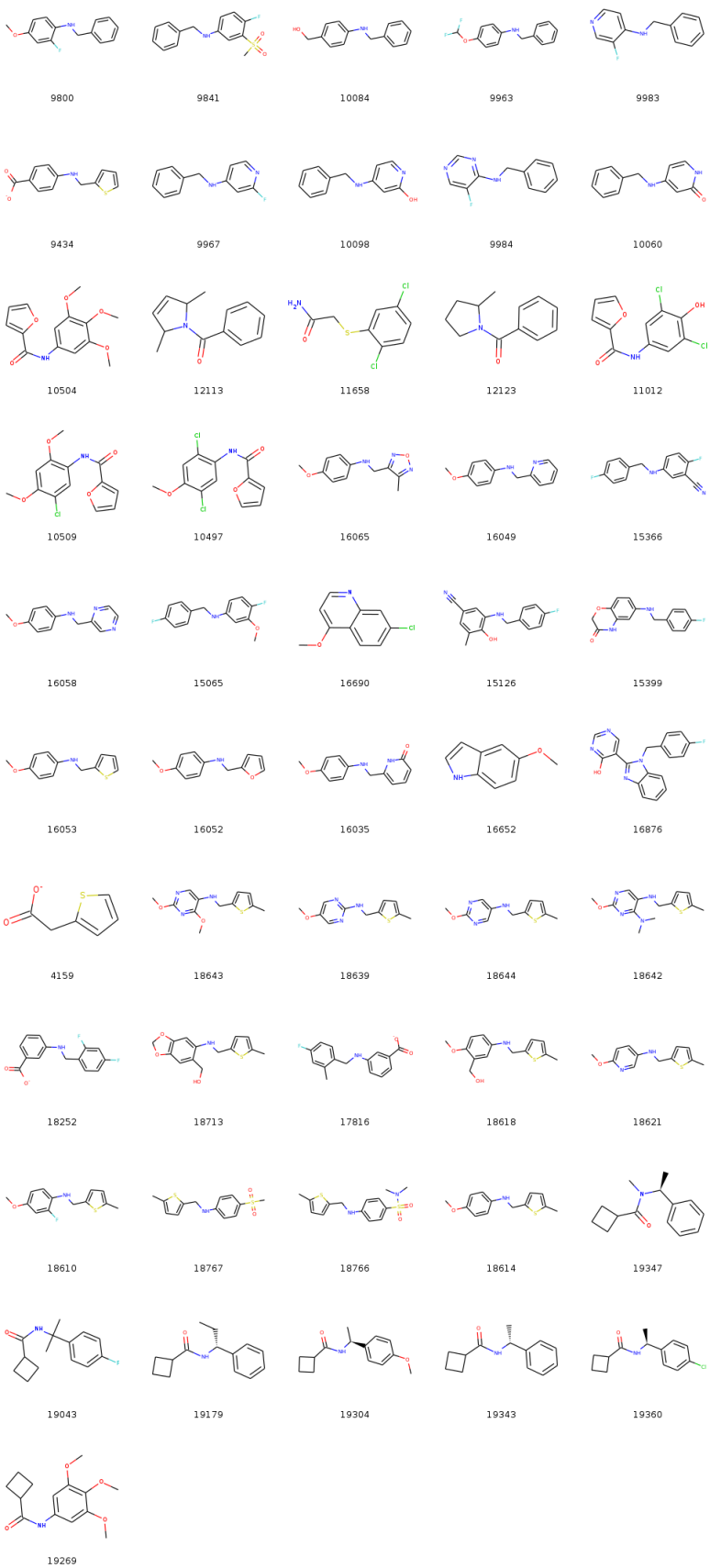



Figure A.5.1 100 molecule subset proposed for synthesis by selection subset 2, maximising fragment diversity.

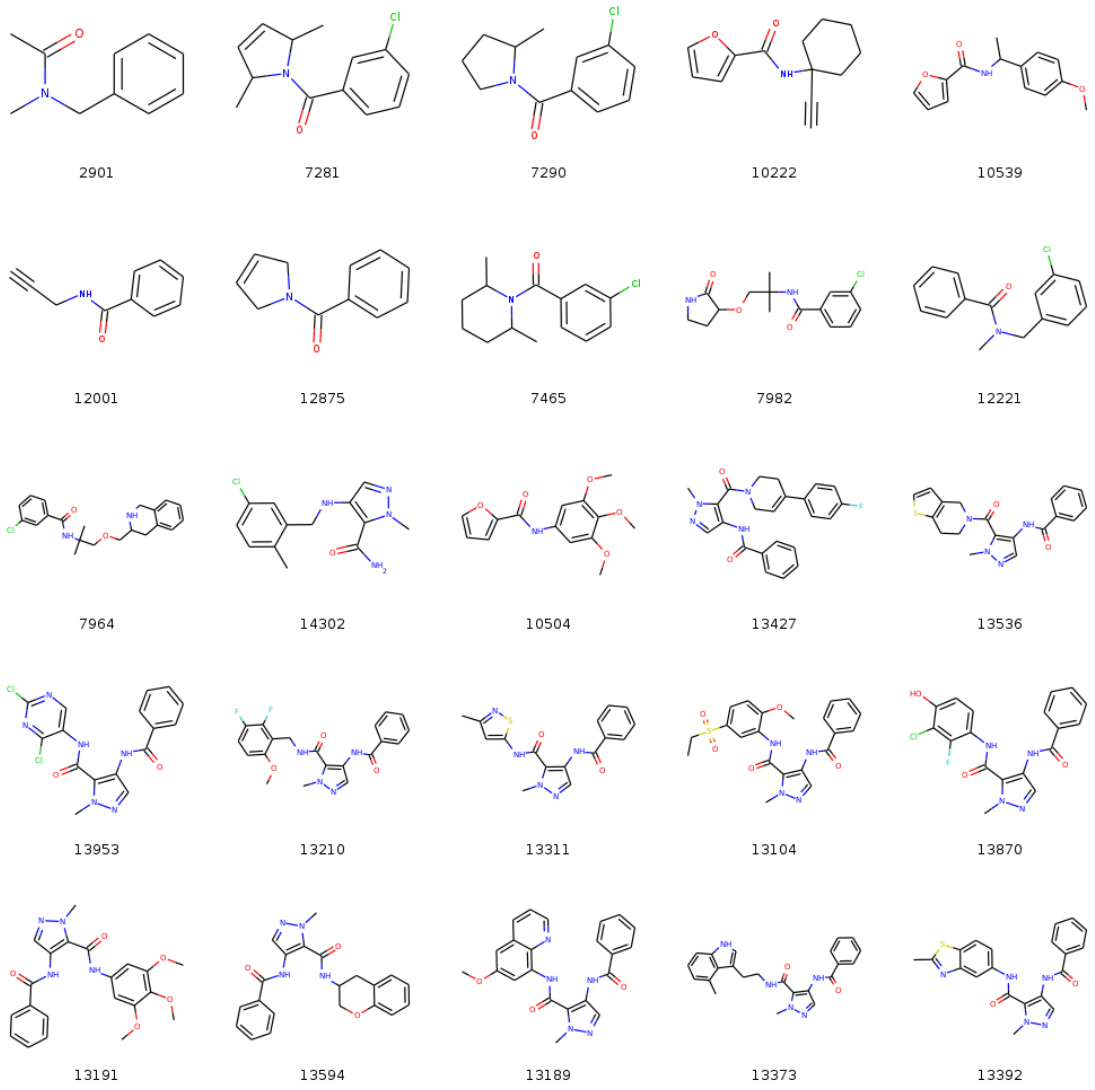



Figure A.5.2 25 molecule subset proposed for synthesis by selection subset 2 to maximise both fragment and PLIF diversity.

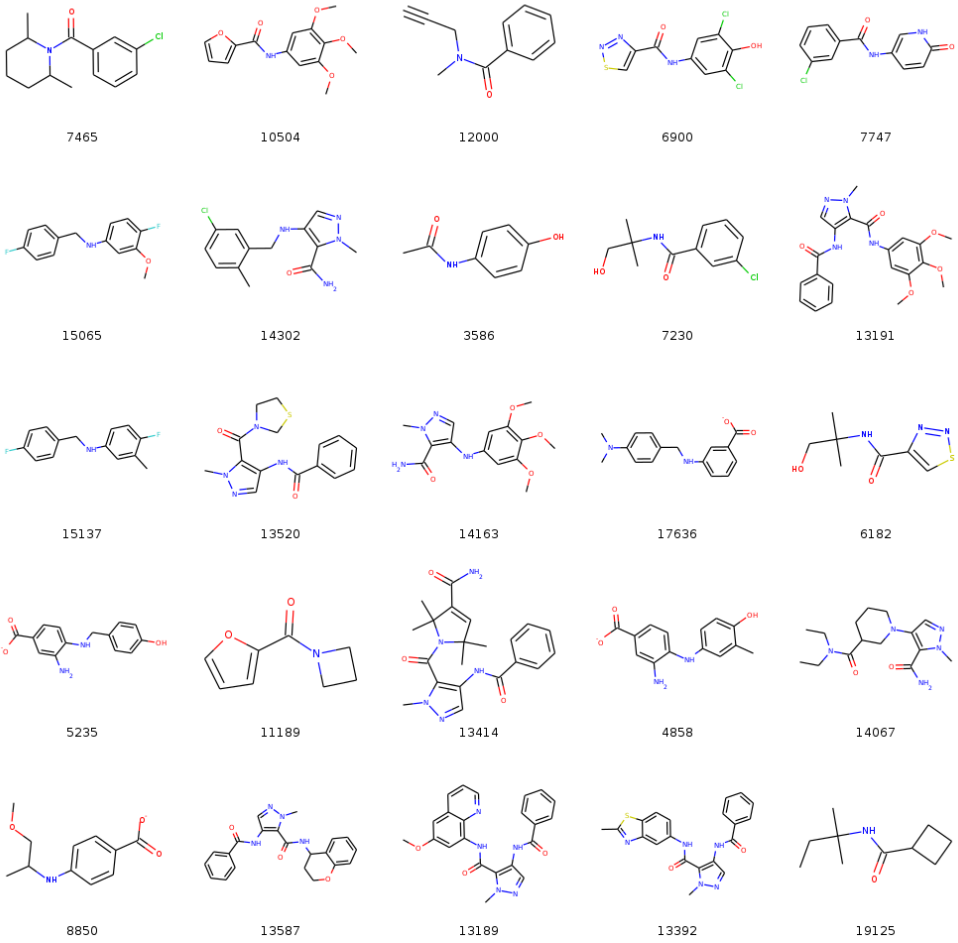



Figure A.5.3 25 molecule subset proposed for synthesis by selection subset 3.2 to minimise the Interaction Score and molecular diversity. The compound ID is shown below each compound.

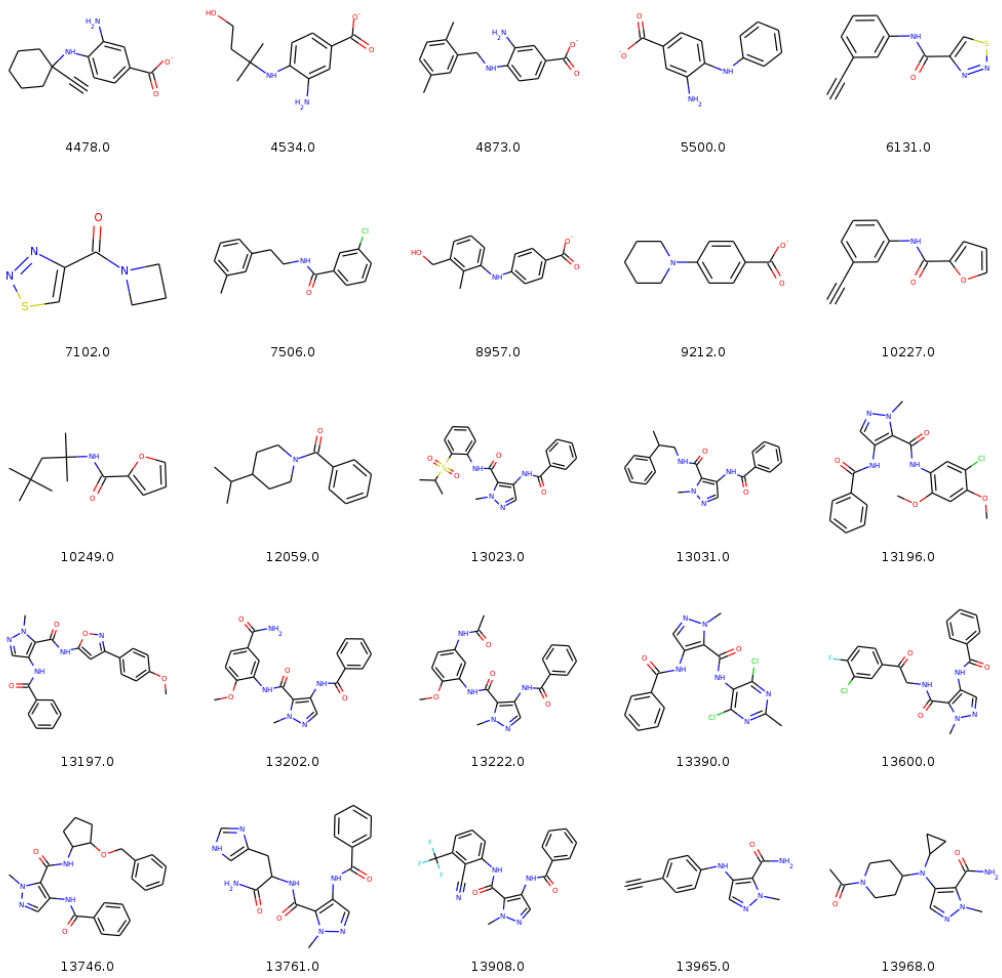



Figure A.5.4 25 molecule subset proposed for synthesis by selection subset 3.2 to maximise the Interaction Score (No Frag) and molecular diversity. The compound ID is shown below each compound.

# Appendix B

# Tables

## Chapter 2

| PDB Code | Ligand Residue Code | Resolution / Å |
|---|---|---|
| 1ajv | NMB | 2.0 |
| 1ajx | AH1 | 2.0 |
| 1c70 | L75 | 2.5 |
| 1d4h | BEH | 1.8 |
| 1d4i | BEG | 1.8 |
| 1d4j | MSC | 1.8 |
| 1dif | A85 | 1.7 |
| 1ebw | BEI | 1.8 |
| 1ebz | BEC | 2.0 |
| 1ec0 | BED | 1.8 |
| 1ec1 | BEE | 2.1 |
| 1ec2 | BEJ | 2.0 |
| 1ec3 | MS3 | 1.8 |
| 1npa | 3NH | 2.0 |
| 1npw | LGZ | 2.0 |
| 1g2k | NM1 | 2.0 |
| 1g35 | AHF | 1.8 |
| 1gno | U0E | 2.3 |
| 1hbv | GAN | 2.3 |
| 1hih | C20 | 2.2 |
| 1hos | PHP | 2.3 |
| 1hps | RUN | 2.3 |
| 1hpv | 478 | 1.9 |
| 1hpx | KNI | 2.0 |
| 1hsg | MK1 | 2.0 |
| 1hte | G23 | 2.8 |
| 1htf | G26 | 2.2 |
| 1htg | G37 | 2.0 |
| 1hvi | A77 | 1.8 |
| 1hvj | A78 | 2.0 |
| 1hvk | A79 | 1.8 |
| 1hvl | A76 | 1.8 |
| 1mob | 0ZQ | 2.5 |
| 1mui | AB1 | 2.8 |
| 1ohr | 1UN | 2.1 |
| 1yt9 | 0IS | 3.0 |
| 1zsf | 0ZS | 2.0 |
| 1zsr | 0ZT | 2.1 |
| 2bbb | HH1 | 1.7 |
| 2bpv | 1IN | 1.9 |
| 2bpx | MK1 | 2.8 |
| 2bpy | 3IN | 1.9 |
| 2bqv | A1A | 2.1 |
| 2cej | 1AH | 2.5 |
| 2cem | 2AH | 1.8 |
| 2cen | 4AH | 1.7 |
| 2pqz | G0G | 1.6 |
| 2pwc | G3G | 1.8 |
| 2pwr | G4G | 1.5 |
| 2qnn | QN1 | 1.5 |
| 2qnp | QN2 | 1.4 |
| 2qnq | QN3 | 2.3 |
| 2upj | U02 | 3.0 |
| 2uxz | HI1 | 1.8 |
| 2uyo | HV1 | 1.8 |
| 2wkz | 5AH | 1.7 |
| 3bgb | LJG | 1.9 |
| 3bgc | LJH | 1.8 |
| 7upj | 1NU | 2.0 |

Table B.2.1 PDB structures used to generate CRANkS grids for Data Set 1.

Smiles String

Cc1ccc(S(=O)(=O)N(CC(C)C)C(CCCCNC(=O)C=Cc2ccccc2)C(=O)O)cc1

COCCCCCN1C(=O)N(CCCCCOC)C(Cc2ccccc2)C(O)C(O)C1Cc1ccccc1

CN(C)CCCCN1C(=O)N(CCCCN(C)C)C(Cc2ccccc2)C(O)C(O)C1Cc1ccccc1

O=C1N(CCCCCO)C(Cc2ccccc2)C(O)C(O)C(Cc2ccccc2)N1CC1CC1

O=C1N(CCCCCn2ccnc2)C(Cc2ccccc2)C(O)C(O)C(Cc2ccccc2)N1CCCCCn1ccnc1

CNC(=S)NCCCCCN1C(=O)N(CCCCCNC(=S)NC)C(Cc2ccccc2)C(O)C(O)C1Cc1ccccc1

O=C(Cc1ccccc1)N(CCCNCCCN(Cc1ccccc1)C(=O)Cc1ccccc1)Cc1ccccc1

O=C1N(CCCCCCO)C(Cc2ccccc2)C(O)C(O)C(Cc2ccccc2)N1CCCCCCO

Cc1ccccc(CN2C(=O)N(Cc3cccc(C)c3)C(Cc3ccccc3)C(O)C(O)C2Cc2ccccc2)c1

CNC(=O)OCCCCCCN1C(=O)N(CCCCCCOC(C)=O)C(Cc2ccccc2)C(O)C(O)C1Cc1ccccc1

CC(C)(C)NC(=O)C1CC2CCCCC2CN1CC(O)C(Cc1ccccc1)NC(=O)C(CC(N)=O)NC(=O)c1ccc2ccccc2n1

Cc1ccc(S(=O)(=O)NCCCCCN2C(=O)N(CCCCCNS(=O)(=O)c3ccc(C)cc3)C(Cc3ccccc3)C(O)C(O)C2Cc2ccccc2)cc1

CC(=O)c1ccccc(CN2C(=O)N(Cc3cccc(C(C)=O)c3)C(Cc3ccccc3)C(O)C(O)C2Cc2ccccc2)c1

CNC(=O)NCCCCCN1C(=O)N(CCCCCNC(=O)NC)C(Cc2ccccc2)C(O)C(O)C1Cc1ccccc1

CS(=O)CCCCCN1C(=O)N(CCCCCO)C(Cc2ccccc2)C(O)C(O)C1Cc1ccccc1

CC(C)CN(CCNCCN(CC(C)C)S(=O)(=O)c1ccccc1)S(=O)(=O)c1ccccc1

Nc1ccc(S(=O)(=O)N(CCCNCCN(Cc2ccccc2)S(=O)(=O)c2ccc(N)cc2)Cc2ccccc2)cc1

CC(C)CN(C(CO)CCCCNC(=O)C(NC(=O)c1cccc(O)c1)C(c1ccccc1)c1ccccc1)S(=O)(=O)c1ccc(N)cc1

CCCCN1C(=O)N(CC2CCCC2)C(Cc2ccccc2)C(O)C(O)C1Cc1ccccc1

Cc1ccc(S(=O)(=O)N(CC(C)C)C(CO)CCCCNC(=O)CN(Cc2ccccc2[N+](=O)[O-])c2ccccc2)cc1

CC(C)CN(C(CO)CCCCNC(=O)CN(Cc1cccc(F)c1)c1ccccc1)S(=O)(=O)c1ccc(N)cc1

CNS(=O)(=O)OCCCCCCN1C(=O)N(CCCCCCO)C(Cc2ccccc2)C(O)C(O)C1Cc1ccccc1

O=C1N(CC=Cc2ccccc2)C(Cc2ccccc2)C(O)C(O)C(Cc2ccccc2)N1CC=Cc1ccccc1

CC(C)CC(NC(=O)C(Cc1ccccc1)CC(O)C(Cc1ccccc1)NC(=O)OC(C)(C)C)C(=O)NC(Cc1ccccc1)C(N)=O

COC(=O)C(CCCCNC(=O)OCC1c2ccccc2-c2ccccc21)N(CC(C)C)S(=O)(=O)c1ccc(C)cc1

NC(CCCN1C(=O)N(CCCC(N)c2ccc(F)cc2)C(Cc2ccccc2)C(O)C(O)C1Cc1ccccc1)c1ccc(F)cc1

COc1ccc(CN(CC(=O)NCCCCC(CO)N(CC(C)C)S(=O)(=O)c2ccc(N)cc2)c2ccccc2)cc1OC

N=C(NO)c1ccccc(CN2C(=O)N(Cc3cccc(C(=N)NO)c3)C(Cc3ccccc3)C(O)C(O)C2Cc2ccccc2)c1

Cc1ccc(S(=O)(=O)N(CC(C)C)C(CCCCNC(=O)NCc2ccccn2)C(=O)O)cc1

O=C1N(Cc2ccccc2F)C(Cc2ccccc2)C(O)C(O)C(Cc2ccccc2)N1Cc1ccccc1F

COCCCCCN1C(=O)N(CCCCCO)C(Cc2ccccc2)C(O)C(O)C1Cc1ccccc1

Cc1ccc(S(=O)(=O)N(CC(C)C)C(CCCCNC(=O)N(C(C)C)C(C)C)C(=O)O)cc1

O=C1N(Cc2cccc(Cl)c2)C(Cc2ccccc2)C(O)C(O)C(Cc2ccccc2)N1Cc1cccc(Cl)c1

C[NH+](C)CCN1C(=O)N(CC[NH+](C)C)C(Cc2ccccc2)C(O)C(O)C1Cc1ccccc1

CNC(=O)CCCCCN1C(=O)N(CCCCCC(=O)NC)C(Cc2ccccc2)C(O)C(O)C1Cc1ccccc1

O=C(NCCCCCN1C(=O)N(CCCCCNC(=O)Nc2ccccc2)C(Cc2ccccc2)C(O)C(O)C1Cc1ccccc1)Nc1ccccc1

CC(C)CN(C(CO)CCCCNC(=O)N(Cc1ccccc1)Cc1ccc(F)cc1)S(=O)(=O)c1ccc(N)cc1

Cc1ccc(S(=O)(=O)N(CC(C)C)C(CCCCNC(=O)C(N)Cc2ccccc2)C(=O)O)cc1

Smiles String

O=C(NCCCCCN1C(=O)N(CCCCCNC(=O)c2ccc(F)cc2)C(Cc2ccccc2)C(O)C(O)C1Cc1ccccc1)c1ccc(F)cc1

CCC(O)CN1C(=O)N(CC(O)CC)C(Cc2ccccc2)C(O)C(O)C1Cc1ccccc1

CC(C)CN(C(CO)CCCCNC(=O)N(Cc1ccc([N+](=O)[O-])cc1)Cc1ccc2c(c1)OCO2)S(=O)(=O)c1ccc(N)cc1

O=C1N(Cc2cccnc2)C(Cc2ccccc2)C(O)C(O)C(Cc2ccccc2)N1Cc1cccnc1

O=C1N(Cc2ccc(CO)cc2)C(Cc2ccccc2)C(O)C(O)C(Cc2ccccc2)N1Cc1ccc(CO)cc1

Cc1ccc(S(=O)(=O)N(CC(C)C)C(CCCCNC(=O)OCC2c3ccccc3-c3ccccc32)C(N)=S)cc1

O=C(Nc1nccs1)c1cccc(CN2C(=O)N(Cc3cccc(C(=O)Nc4nccs4)c3)C(Cc3ccccc3)C(O)C(O)C2Cc2ccccc2)c1

O=C1N(Cc2cccc(-c3nnn[nH]3)c2)C(Cc2ccccc2)C(O)C(O)C(Cc2ccccc2)N1Cc1cccc(-c2nnn[nH]2)c1

Cc1ccc(S(=O)(=O)N(CC(C)C)C(CCCCNC(=O)c2ccccc2)C(=O)O)cc1

CCCC(N=O)c1cccc(CN2C(=O)N(Cc3cccc(C(CCC)N=O)c3)C(Cc3ccccc3)C(O)C(O)C2Cc2ccccc2)c1

CC(C)CCN(C(CO)CCCCNC(=O)N(Cc1ccc2c(c1)OCO2)Cc1cccs1)S(=O)(=O)c1ccc(N)cc1

COc1ccc(CN(Cc2ccc3c(c2)OCO3)C(=O)NCCCCC(CO)N(CCC(C)C)S(=O)(=O)c2ccc(N)cc2)cc1

CCNC(=O)c1cccc(CN2C(=O)N(Cc3cccc(C(=O)NCC)c3)C(Cc3ccccc3)C(O)C(O)C2Cc2ccccc2)c1

O=S(=O)(c1ccccc1)N(CCCNCCCN(Cc1ccccc1)S(=O)(=O)c1ccccc1)Cc1ccccc1

O=C(NCCCCCN1C(=O)N(CCCCCNC(=O)Oc2ccc([N+](=O)[O-])cc2)C(Cc2ccccc2)C(O)C(O)C1Cc1ccccc1)OCc1ccncc1

CC(=CCN1C(=O)N(CC=C(C)C(=O)O)C(Cc2ccccc2)C(O)C(O)C1Cc1ccccc1)C(=O)O

CC(C)CN(C(CO)CCCCNC(=O)C(NC(=O)c1cccnc1)C(c1ccccc1)c1ccccc1)S(=O)(=O)c1ccc(N)cc1

O=C1N(Cc2cccc(O)c2)C(Cc2ccccc2)C(O)C(O)C(Cc2ccccc2)N1Cc1ccc2[nH]ncc2c1

OC(C(COCc1ccccc(F)c1)OCc1ccccc1)C(O)C(COCc1ccccc(F)c1)OCc1ccccc1

CC(C)(C)OC(=O)NCCCCC(NC(=O)OCc1ccccc1)C(=O)O

Cc1cc(C)cc(NCC=CCN2C(=O)N(CC=CCNc3cc(C)cc(C)c3)C(Cc3ccccc3)C(O)C(O)C2Cc2ccccc2)c1

C=COCCN1C(=O)N(CCOC=C)C(Cc2ccccc2)C(O)C(O)C1Cc1ccccc1

Cc1ccc(S(=O)(=O)NCCCCC(NC(=O)OCC2c3ccccc3-c3ccccc32)C(=O)O)cc1

COCCOCn1cc(C=CCN2C(=O)N(CC=Cc3cnn(COCCOC)c3)C(Cc3ccccc3)C(O)C(O)C2Cc2ccccc2)cn1

Cc1ccc(COC(COCc2ccccc2)C(O)C(O)C(COCc2ccccc2)OCc2ccc(C)cc2)cc1

Cc1ccc(S(=O)(=O)N(Cc2ccccc2)C(CCCCNC(=O)OCC2c3ccccc3-c3ccccc32)C(=O)O)cc1

O=C1N(Cc2ccc3ccccc3c2)C(Cc2ccccc2)C(O)C(O)C(Cc2ccccc2)N1CC1CC1

O=C1N(Cc2ccc(CO)cc2)C(Cc2ccccc2)C(O)C(O)C(Cc2ccccc2)N1Cc1ccc(CO)cc1

Cc1ccccc1COC(COCc1ccccc1F)C(O)C(O)C(COCc1ccccc1F)OCc1ccccc1C

Cc1ccc(S(=O)(=O)N(CC(C)C)C(CO)CCCCNC(=O)CN(CCc2ccccc2)c2ccccc2)cc1

O=C1N(Cc2cccc(-c3ncc[nH]3)c2)C(Cc2ccccc2)C(O)C(O)C(Cc2ccccc2)N1Cc1cccc(-c2ncc[nH]2)c1

Cc1ccc(S(=O)(=O)N(CC2CCCC2)C(CCCCNC(=O)OCC2c3ccccc3-c3ccccc32)C(=O)O)cc1

CC(C)CN(C(CCCCNC(=O)OCC1c2ccccc2-c2ccccc21)C(=O)O)S(=O)(=O)c1cccc2ccccc12

CC(C)CN(C(CO)CCCCNC(=O)N(Cc1ccccc1)Cc1ccncc1)S(=O)(=O)c1ccc(N)cc1

OC(C(COCc1ccccc1)OCc1ccccc1)C(O)C(COCc1ccccc1)OCc1ccccc1

O=S(=O)(c1ccccc1)N(CCNCCN(Cc1ccccc1)S(=O)(=O)c1ccccc1)Cc1ccccc1

OC(C(COc1ccccc1)OCc1ccccc1)C(O)C(COc1ccccc1)OCc1ccccc1

O=C(NC(Cc1ccccc1)C(O)CC1(Cc2ccccc2)N=CC(C2c3ccccc3CC2O)C1=O)OC1CCOC1

COc1ccc(CN(Cc2ccc3c(c2)OCO3)C(=O)NCCCCC(CO)N(CCC(C)C)S(=O)(=O)c2ccc(N)cc2)cc1OC

Smiles String

Cc1ccc(S(=O)(=O)N(CC(C)C)C(CNC(=O)OCC2c3ccccc3-c3ccccc32)C(=O)O)cc1

OC(C(COCc1ccccc1)OCc1ccc(F)cc1)C(O)C(COCc1ccccc1)OCc1ccc(F)cc1

COc1ccc(CN(CC(=O)NCCCCC(CO)N(CC(C)C)S(=O)(=O)c2ccc(N)cc2)c2ccccc2)cc1

OC(C(COCc1ccccc1)OCc1ccccc1F)C(O)C(COCc1ccccc1)OCc1ccccc1F

Cc1ccc(S(=O)(=O)N(CC(C)C)C(CCCCNC(=O)OCC2c3ccccc3-c3ccccc32)c2nnn[nH]2)cc1

C=CCN1C(=O)N(CCCC)C(Cc2ccccc2)C(O)C(O)C1Cc1ccccc1

CNC(=O)C(NC(=O)C(OCc1ccccc1)C(O)C(O)C(OCc1ccccc1)C(=O)NC(C(=O)NC)C(C)C)C(C)C

OCc1ccc(COC(COCc2ccccc2)C(O)C(O)C(COCc2ccccc2)OCc2ccc(CO)cc2)cc1

CC(C)CN(C(CO)CCCCNC(=O)N(Cc1ccc2c(c1)OCO2)Cc1ccc2c(c1)OCO2)S(=O)(=O)c1ccc(N)cc1

CC(C)CN(C(CCCCNC(=O)OCC1c2ccccc2-c2ccccc21)C(=O)O)S(=O)(=O)c1ccc(N)cc1

COc1cc2c(Nc3ccc(Br)cc3F)ncnc2cc1OCC1CCN(C)CC1

CNC(=O)C(NC(=O)C(OCc1ccc(CCc2ccccc2)cc1)C(O)C(O)C(OCc1ccc(CCc2ccccc2)cc1)C(=O)NC(C(=O)NC)C(C)C)C(C)C

COC(=O)NC(Cc1ccc2ccccc2c1)C(=O)NCCCCC(CO)N(CC(C)C)S(=O)(=O)c1ccc(N)cc1

O=C1N(CC(O)Cc2ccccc2)C(Cc2ccccc2)C(O)C(O)C(Cc2ccccc2)N1CC(O)Cc1ccccc1

OC(C(COCc1ccc(Br)cc1)OCc1ccccc1)C(O)C(COCc1ccc(Br)cc1)OCc1ccccc1

CC(C)c1nc(CN(C)C(=O)NC(C(=O)NC(Cc2ccccc2)CC(O)C(Cc2ccccc2)NC(=O)OCc2cncs2)C(C)C)cs1

CC(C)CN(C(CO)CCCCNC(=O)C(NC(=O)C1CC1)C(c1ccccc1)c1ccccc1)S(=O)(=O)c1ccc(N)cc1

N#Cc1ccc(COCC(OCc2ccccc2)C(O)C(O)C(COCc2ccc(C#N)cc2)OCc2ccccc2)cc1

CCCCN1C(=O)N(CCCC)C(Cc2ccccc2)C(O)C(O)C1Cc1ccccc1

COCCOCCN1C(=O)N(CCOCCOC)C(Cc2ccccc2)C(O)C(O)C1Cc1ccccc1

O=C1N(CC2CC2(c2ccccc2)c2ccccc2)C(Cc2ccccc2)C(O)C(O)C(Cc2ccccc2)N1CC1CC1(c1ccccc1)c1ccccc1

COc1ccccc1CN1C(=O)N(Cc2ccccc2OC)C(Cc2ccccc2)C(O)C(O)C1Cc1ccccc1

CC(C)CN(C(CO)CCCCNC(=O)C(NC(=O)N1CCOCC1)C(c1ccccc1)c1ccccc1)S(=O)(=O)c1ccc(N)cc1

O=C1N(CC=Cc2ccccc2O)C(Cc2ccccc2)C(O)C(O)C(Cc2ccccc2)N1CC=Cc1ccccc1O

O=C1N(Cc2ccc(CO)cc2)C(Cc2ccccc2)C(O)C(O)C(Cc2ccccc2)N1Cc1ccc2ccccc2c1

CC(C)(C)C(=O)c1cccc(CN2C(=O)N(Cc3ccccc(C(=O)C(C)(C)C)c3)C(Cc3ccccc3)C(O)C(O)C2Cc2ccccc2)c1

O=C1N(CCCCCO)C(Cc2ccccc2)C(O)C(O)C(Cc2ccccc2)N1CCCCCCO

COC(=O)NC(CCc1ccccc1)C(=O)NCCCCC(CO)N(CC(C)C)S(=O)(=O)c1ccc(N)cc1

Cc1ccc(S(=O)(=O)N(CC(C)C)C(CCCCNC(=O)CN(C)c2ccccc2)C(=O)O)cc1

CC(C)CN(C(CO)CCCCNC(=O)C(NC(=O)c1nccccO)C(c1ccccc1)c1ccccc1)S(=O)(=O)c1ccc(N)cc1

Cc1ccc(S(=O)(=O)NC(CCCCNC(=O)OCC2c3ccccc3-c3ccccc32)C(=O)O)cc1

CNC(=O)C(NC(=O)C(OCc1ccc(C=CC(=O)OC)cc1)C(O)C(O)C(OCc1ccc(C=CC(=O)OC)cc1)C(=O)NC(C(=O)NC)C(C)C)C(C)C

Cc1ccc(S(=O)(=O)N(CC(C)C)C(CCCCNC(=O)OCc2ccccc2)C(=O)O)cc1

O=C(Cc1ccccc1)N(CCNCCN(Cc1ccccc1)C(=O)Cc1ccccc1)Cc1ccccc1

COc1cc(C(=O)NC(C(=O)NCCCCC(CO)N(CC(C)C)S(=O)(=O)c2ccc(N)cc2)C(c2ccccc2)c2ccccc2)cccIO

CC(C)CN(C(CO)CCCCNC(=O)C(NC(=O)OCc1ccccn1)C(c1ccccc1)c1ccccc1)S(=O)(=O)c1ccc(N)cc1

COC(=O)NC(C(=O)NCCCCC(CO)N(CC(C)C)S(=O)(=O)c1ccc(N)cc1)C(c1ccccc1)c1ccccc1

Cc1ccc(S(=O)(=O)N(CC(C)C)C(CCCCNC(=O)CSc2ccccc2)C(=O)O)cc1

| Smiles String |
| --- |
| Cc1ccc(COCC(OCc2ccc(C)cc2)C(O)C(O)C(COCc2ccc(C)cc2)OCc2ccc(C)cc2)cc1 |

Table B.2.2 SMILES of active compounds used as candidate compounds in Data Set 1.

| Smiles String |
| --- |
| CCOc1ccc(C2C(C(=O)c3ccc(C)cc3)=C([O-])C(=O)N2c2nc(C)c(C(C)=O)s2)cc1OCC |
| N#CC=C[NH+]1CCCC1 |
| OCCC(F)(F)F |
| Clc1cccc(C=NNc2ccc(I)cc2)c1Cl |
| CC1(C)CCO1 |
| C1=NCN=C2N=CC=C12 |
| O=C=NCCl |
| O=C1CSCCN1 |
| CC1CCC(C(C)C)C(OP(=S)(OC2CC(C)CCC2C(C)C)OC2CC(C)CCC2C(C)C)C1 |
| Nc1nn(-c2cncc3nnnn23)cc1Br |
| S=C(NC1CCCCCCC1)N(CCCN1CCOCC1)CC1CC=CCC1 |
| CC(C)(C)c1ccc(C2=Nc3ccc(F)cc3C(c3ccccc3)=NN2)cc1 |
| CCCCC(=O)C(C(O)C(C)CC=CCO)N(C)C(=O)C(NC(=O)C(CC(C)C)[NH2+]C)C(C)C |
| CSc1nc(NC(C)C)c2c3c(sc2n1)CCCC3 |
| CON=CNOC |
| CCC1=C(C)C(Br)=NC1=Cc1[nH]c(Br)c(CC)c1C |
| O=C(NCCSSCCNC(=O)C12CC3CC(CC(C3)C1)C2)C12CC3CC(CC(C3)C1)C2 |
| C#CCCC1CO1 |
| C=C=CN(C)C |
| O=S(=O)([N-]c1ccc(OC(F)F)cc1)c1sccc1Br |
| CC(C)N=C1N(C(C)C)C(=O)C1(C(=O)C(C)(C)C)C1(C(C)(C)C)OC(=O)C(C(=O)C(C)(C)C)=C(C(C)(C)C)O1 |
| OCc1cc(CCN(O)CCCCCCCCN(O)CCc2ccc(O)c(CO)c2)ccc1O |
| CC1=NN=CC1Cl |
| O=[N+]([O-])c1cc(Cl)cc(C(c2cc(Cl)cc([N+](=O)[O-])c2[O-])C(Cl)(Cl)Cl)c1[O-] |
| N#Cc1ncccc1CNC1CCc2ccccc2C1 |
| COC1C(O)C(C)OC(OC2CCC3(C)C(CCC4C3CCC3(C)C(C5=CC(=O)OC5)CCC43O)C2)C1O |
| NC(=S)NN=C1CCS(=O)(=O)CC1 |
| C=CC(C)(O)CCC(C)=CCCC(O)C(C)(C)OC1OC(CO)C(O)C(O)C1O |
| NC(=C[O-])C(O)C(O)C(O)COP(=O)([O-])[O-] |
| O=C1NC(=Nc2ccc(Br)cc2)SC1=Cc1ccc(SC2N=C3C=CC=CC3N2)o1 |
| O=C(NCCCn1c(=O)[nH]c2ccccc21)c1csc(Nc2ccccc2)[n+]1Cc1cccs1 |
| NC1=CC(=O)C(=C2N=C(c3ccncc3)[N-]O2)C=C1 |
| CN(C1CCSC1)C(C[NH3+])c1cc(Br)c2c(c1)OCO2 |
| CCSC(C)(C)C |

Smiles String

O=C([O-])c1cc(S(=O)(=O)N2CCCCC2)ccc1Sc1ccc2c(c1)CCC2

O=P1(S(=O)(=O)Oc2ccccc2)N[NH+]2CC3C=C(c4ccccc4O)C(O)=CC3CC2CO1

CNC(C)=[NH+]C#N

CC(c1nnc(SCc2nc(N)nc(Nc3ccc(F)cc3)n2)n1-c1ccc(Cl)cc1)[NH+](C)C

OC1CCC2OC12

CCN1NC(C)=C(Br)C1C(=O)N1CC(C(=O)NC(C)C2=C(C)N=NC2C)C2(CC[NH2+]CC2)C1

CSC1=NS(=O)(=O)c2cc(N(C)C)ccc2N1

C=C1CN1C

COc1ccc2c(c1)[nH]c1c(C([O-])=Nc3ccc(Cl)cc3)[nH+]ccc12

N#CC(=Cc1cn(-c2ccccc2)nc1-c1cc2ccccc2o1)c1nc([O-])c2ccccc2n1

CC1(C)CC2=CC3(C)CC(C)(C)CC4(C3)NC(=O)NC(=C24)C1

O=c1[nH]c([O-])nnc1SCCOc1c(Cl)c(Cl)c(Cl)c(Cl)c1Cl

O=C1[NH+]=c2c(Br)cc(Br)cc2=C(O)C1C1[NH+]=c2ccccc2=[NH+]1

COc1ccc(C=C2CNCC3=C2OC(N)=C(C#N)C3c2ccc(OC)c(OC)c2OC)c(OC)c1OC

CC1=C(C2=C(C)C(C(C)C)C3=C(C(=O)C(O)C3O)C2O)C(O)C2=C(C(O)C(O)C2=O)C1C(C)C

c1nc(N2CCc3[nH]ncc3C2)c2ccsc2n1

C=C(C)CC1C=CCC(CC(=C)C)[NH+]1CC([O-])C(F)(F)C(F)(F)C(F)(F)F

CC(C)N(CC(=O)N1c2ccccc2-n2cccc2C1c1ccco1)C(=O)Nc1ccccc1F

N#CC1C(=O)N=C(N2CCCC2)NC12CCCC2

[NH-]S(=O)(=O)N=C(N)CCS(=O)Cc1csc(N=C(N)N)n1

Cc1noc(C)c1CSCC(=O)NN

Nn1c([S-])nnc1-c1cccc2c1OCCO2

CCn1cncn1

N=C1C=CC(c2nc(-c3cc(F)c(F)c(F)c3)no2)S1

CCCCOc1ccc(-n2sc3c(c2=S)-c2ccccc2NC3(C)C)cc1

O=c1[nH]cc(S(=O)(=O)[N-]C2=NC3=CCC=CC3=N2)c(=O)[nH]1

O=C([O-])C1CCCCC1C(=O)CCCCOC(=O)C1CCCCC1C(=O)[O-]

Cc1ccc(NS(=O)(=O)c2ccc3c(c2)C2C=CCC2C(c2cccc([N+](=O)[O-])c2)N3)c(C)c1

Nc1nc2cc(Br)ccc2n1C1CC[NH+]2CCCCC12

CC1(C)Cc2c(sc(NC(=O)c3cc(Cl)sc3Cl)c2C#N)C(C)(C)[NH2+]1

Cc1c(C)c(C)c(S(=O)(=O)NC2CCN(CC(F)(F)F)C2)c(C)c1C

Clc1cc(Cl)c(Nc2nsnc2Cl)cc1Cl

N#CCCP

COc1ccc(C(O)(c2ccc(OC)cc2)c2ccc(C[NH+](Cc3ccsc3)CC3CCCO3)o2)cc1

Cc1ccc(C=Nc2ccc(Cc3ccc(N=Cc4ccc(C)s4)cc3)cc2)s1

CC(C)=CCCC1(C)C=Cc2c(O)c3c(c(CC=C(C)C)c2O1)OC12C(=CCCC1C(C)(C)OC2(CC=C(C)C(=O)[O-])C(=O)[O-])C3=O

O=C(COC(=O)c1cc(-c2ccc(Br)cc2)n[nH]1)NCC(=O)Nc1ccc(F)c(F)c1

CC1=C(C)C(Cc2[nH]c(CC3=NC(C(=O)OC(C)(C)C)C(C)=C3C)c(C)c2C)[NH+]=C1C(=O)OC(C)(C)C

C#CCOCC(O)C[NH+](CCOC)Cc1c(CC)nn(-c2ccccc2)c1OC1ccccc1OC

CNc1cc(-c2cccnc2Nc2c(C)ccc3c(Nc4ccc5sc(C)nc5c4)nccc23)ncn1

Smiles String

CC(=O)C(F)F

NC(=O)C1CCN(C(=O)C2CC(O)C(O)C3NC(=S)N(c4ccc(-n5cccn5)cc4)C23)CC1

CC1(CO)OC(C)(Oc2c[nH]c3ccc(I)cc23)C(C)(O)C(C)(O)C1(C)O

CCCn1c(N)c(N(CCOC)C(=O)C(F)(F)F)c(=O)n(CC(=O)NC(CC)CC)c1=O

O=C([O-])COc1ccc(C=NNC(=O)C(=CC=Cc2ccccc2)NC(=O)c2ccccc2)cc1

N#CC1CC2C=CC1C2

Cc1cc(C)cc(NCc2ccc3[nH+]c(NC4CC[NH+](CCO)CC4)n(Cc4[nH+]c(C)ccc4[O-])c3c2)c1

Nc1ncnc2nc(-c3ccc(N4CCOCC4)nc3)cc(NCCC3=c4ccccc4=[NH+]C3)c12

CCOC1(c2cccs2)OC(c2cccs2)=C(c2cccs2)OC1c1cccs1

CC12CCC(c3c(C(F)(F)F)nc([S-])nc31)C2(C)C

O=C(CSc1nnc(-c2ccc[nH]2)o1)N1N=C2C(=Cc3ccccc3)CCCC2C1c1ccccc1

CC1CN(C(=O)C(C)C(N)=NO)C(C)CO1

CC(Nc1ncnc2[nH]cnc12)c1nncn1C

CCCC(CC)COc1cc(-c2nnc(-c3ccc(N)cc3)o2)cc(-c2nnc(-c3ccc(N)cc3)o2)c1

C1=NCCCCCC1

COc1cc(-c2[nH+]c3ccccn3c2NC(C)(C)CC(C)(C)C)ccc1OC(=O)C(C)(C)CCl

O=C1C2C=CSC2NCN1CC1N=NC2=C1CCC2

C[NH+]1CCc2c(sc3c2C(=O)NC(c2ccc(-c4ccc(F)cc4)o2)N3)C1

C[NH2+]C1(P(=O)([O-])O)CC(C)(P(=O)([O-])O)P(=O)([O-])OP1(=O)[O-]

Cc1cc(C)c2c(CC(=O)NN=C3CCCc4oc(C(=O)Nc5ccc6c(c5)OCO6)c(C)c43)coc2c1

CC(C)n1c(N2CCCC3CCCCC32)n[nH]c1=S

Cc1nn(C)c(C)c1CN(C)c1n[nH]c(=O)[nH]c1=O

CC1=Nc2ccccc2C1C(c1ccccn1)N1CCN(C(c2ccccn2)c2c(C)[nH]c3ccccc23)CC1

Oc1ccccc1C1CCC2(CC(O)CN2c2ccccc2O)C1

C=CCSC1=NC(=O)C(C(c2ccc([N+](=O)[O-])cc2)c2c(N)nc(SCC=C)[nH]c2=O)C(N)=N1

Cc1cc(C)nc(NC(=NC(=S)Nc2cccc(C)c2Cl)NCCSCc2nc[nH]c2C)n1

CN(c1cc[nH+]cc1)c1ccc(-c2ccc3c(C(N)=O)nn(-c4ccc5onc(N)c5c4)c3c2F)c(F)c1

O=c1[nH]c2cc(Br)c(C(Cl)c3cn[nH]c3)cc2[nH]1

CCCCCC(O)C=CC1C=CC(=O)C1CC=CCCCC(=O)NCCCCCNC(=O)CCCCC1SCC2NC(=O)NC21

CC(Nc1ccc2c(c1)OC(C)(C)O2)c1ccsc1

[O-][n+]1ccc(CC(c2ccc(NCc3ccccc3)nc2)c2ccc(OC(F)F)c(OC(F)F)c2)cc1

CCC(c1nnnn1C(C)(C)CC)[NH+](CCc1cccc(C)c1)Cc1cc2cc(C)cc(C)c2[nH]c1=O

COc1cc2c(cc1O)CC1c3c(cc(OC)c(O)c3-2)CC[NH+]1C

CCCSc1nnc(N=C([O-])c2cc(Br)c(Br)s2)s1

CCOC(=O)NNC(=S)NP(=O)(NC(=S)NNC(=O)OCC)OCC

CC(=O)OC1CC2C(COC(=O)CC(C)C)=COC(OC(=O)CC(C)C)C2C1(O)COC(=O)CC(C)C

OCC#Cc1csc(CSc2ncccn2)c1

C=CCOc1c(C)cc(C(C)(C)C)c2cc(C)c(OCC=C)c(C)c2)cc1C

Cc1cccc([N-]S(=O)(=O)c2ccc(-c3cnc(C4CCCC4)o3)s2)c1C

CCc1c(C(=O)[O-])nnn1C(C)C(=O)NC(=O)NC

Cc1cc(C2=NC3NC(=O)N=C([O-])C3C2C2C(=O)NC(=O)N=C2N)c(C)s1

Smiles String

CCC(CO)[NH2+]Cc1cc(Br)c(OCc2ccc(F)cc2Cl)c(OC)c1

O=C1N=C(C(=Cc2ccc(C(=O)[O-])cc2)c2cccs2)N=C2SCC(c3ccc(Cl)cc3)=C12

CC1OC1CCCl

C=CC(C)CC(C)O

CCCC1CC(C(=O)NC(C(C)Cl)C2OC(SC)C(O)C(O)C2OP(=O)([O-])[O-])[NH+](C)C1

CCN1C(=O)NC(c2ccc(Cl)cc2Cl)C2=C1CN(CCC(=O)N1CC3(C)CC1CC(C)(C)C3)C2=O

Cc1ccc2c(c1)C1(C(=O)N2)C2(C#N)C(N)=[NH+]C(ON=C3CCCCC3)(ON=C3CCCCC3)C21C#N

Clc1ccccc1-c1ccc(N2N=CN3N=C([N-]c4ccccc4)NC(=Nc4ccccc4)C32)o1

C#CCN1C(=[NH2+])C(c2ccc(F)cc2)[NH+]=C1C1CCCC1

CC1(C)SC2C([NH3+])C(=O)N2C1C(=O)[O-]

CC(C)=C1C2CCC1C(C(=O)NNC(=O)c1csc3c1CCC(C(C)(C)C)C3)C2C(=O)[O-]

O=C1CCCC2C3CCC[N+]4([O-])CCCC(O)(CN12)C34

CC(C)N1N=CC2=NC(c3ccc(Cl)s3)NC21

Cc1cc(Br)cc2c([O-])cc(C(F)(F)F)nc12

CCCCNS(=O)(=O)c1ccc2nc(-c3cccnc3)cc(C(=O)NCC(OCC)OCC)c2c1

CC1(C)CC(OP(OC2CC(C)(C)N(O)C(C)(C)C2)OC2CC(C)(C)N(O)C(C)(C)C2)CC(C)(C)N1O

c1ccc2nc(NN=C3C4CC5CC(C4)CC3C5)ccc2c1

FC(F)(F)CSc1ccccc1NCc1ccc(Br)o1

C=CCn1c(=NC(=O)CCCSc2ccc(F)cc2)sc2cc(F)cc(F)c21

Cc1ccc(SCC(CSCCO)OCn2cnc3c2NC(N)=NC32OCCO2)cc1

Cc1c[nH]c(C)c1C

COc1cc(OC)cc(C(=O)NC(CCSC)C(=O)N2CCC3(CC2)NCCc2[nH+]c[nH]c23)c1

Brc1cccc(C=NN=c2[nH]c3ccccc3s2)c1

Nc1cccc(-n2ccc(C(=O)NCc3nnn[n-]3)n2)c1

C=CCNc1nc2c(s1)CC1C(C)(CO)C(O)CCC1(C)C2CC(=O)NC1CC1

N#CC(C(N)=O)C(c1ccc(Br)cc1)c1c(-c2ccccc2)nn(-c2ncc(-c3ccccc3)s2)c1O

CC(C)CCC1=NC(C2C=C(C(C)C)N=N2)N2CC[NH+](Cc3cn(CC(N)=O)c4ccccc34)CC12

COc1ccc2ncc(F)c(CCN3CCC([NH2+]Cc4cc5c(s4)CCCN5)C(O)C3)c2n1

CC(=NO)C1=CCC2C3CC=C4CC(O)CCC4(C)C3CCC12C

C1=CCC(c2nc(-c3ccccc3)c(-c3ccc(-c4[nH]c(C5CC=CCC5)[nH+]c4-c4ccccc4)cc3)[nH]2)CC1

CC(=CC(=O)[O-])C=C(C)CC(C)CCCCCC(O)C(CO)C(=O)[O-]

O=C(NNC1=CCCCCN1)c1ccncc1

Cc1oc2c(C)c3oc(=O)c(CC(=O)N(C)CC(O)C(O)C(O)C(O)CO)c(C)c3cc2c1C

O=C(C=Cc1ccc(-c2ccc(Cl)cc2Cl)o1)NC(=S)Nc1nc2ccc(Cl)cc2s1

Cc1nn(-c2nc3c(cc2C[NH3+])CCCC3)c(C)c1Br

O=C([O-])CNC(=O)CC1SC([N-]N=C2CSC(=O)N2)=NC1=O

Cc1c(C(=O)ON=C(N)c2cc(C(F)(F)F)cc(C(F)(F)F)c2)sc2cc(C(C)(C)C)sc12

N#CCN(CC([O-])=Nc1c(Cl)cccc1C(F)(F)F)Cc1ccccc1

CCC(C)(C)C(=O)OC1CC(C(=O)[O-])C=C2C=CC(C)C(CCC3CC(O)CC(=O)O3)C21

CC1=NN(C(=O)c2ccoc2C)C(O)(C(F)(F)F)C1

CCn1c(SCC#N)nnc1-c1nonc1N

Smiles String

CCC(C)(C)OCc1ccc(-c2nc(CCl)cs2)cc1

OCCOc1ccccc(C[NH+]2CCOc3ccc(C4(O)CCN(c5ccccc5F)CC4)cc3C2)c1

CCCCNC1=C(C(=O)N2CCOCC2)C(=N)C(C(=O)Nc2sc3c(c2C(=O)OC)CCC3)S1

O=S1OCCCO1

NC(=O)CN=O

COC(=O)CCC(C)C1CCC2C3CCC4C=CCCC4(C)C3C=CC12C

Cc1ccccc1C1C(C=C2SC(=S)N=C2[O-])=C(c2ccccc2)Oc2ccccc21

CC[NH+](CC)CCCC(C)NC(=O)CSc1nc(Cl)cc(N2CCN(C(=O)NC(C)(C)C)C(C)C2)n1

CCCNC1c2c(F)ccc(F)c2S(=O)(=O)C1C

Clc1ccccc1Cc1nc2ccccc2n1CCCOc1ccccc1Cl

CCCCCN(C)S(=O)(=O)c1c(Cl)cc(Br)cc1Cl

FC(F)(F)C(Oc1nc(NC2CCCCC2)nc(N(c2ccccc2)c2ccccc2)n1)C(F)(F)F

COP(=S)(Nc1c(C(C)(C)C)cc(C(C)(C)C)cc1C(C)(C)C)c1cccc2cccc(N(C)C)c12

O=C(NC(N1CCCCC1)C(Cl)(Cl)Cl)NC(N1CCCCC1)C(Cl)(Cl)Cl

Nc1ccc(=NS(=O)(=O)c2ccccc2Br)[nH]c1

CC[NH+]1CC2(OC(=O)c3ccccc3NC(C)=O)CCC(OC)C34C1C(CC23)C1(O)CC(OC)C2CC4C1(O)C2OC

Cc1cccn2c(=O)c(C=Nc3c(C)n(C)n(-c4ccccc4)c3=O)c(Nc3ccc(SC(F)F)cc3)nc12

Cc1cc(C)c(NC(=S)NCC2(C)CC(NC(=S)Nc3c(C)cc(C)cc3C)CC(C)(C)C2)c(C)c1

CCCCN1C(=O)C2C(C(=O)Nc3cccc(SC)c3)C3C=CC2(O3)C1C(=O)NC1CCCC(C)C1C

CC=CC1C(C(=O)c2c([O-])c(-c3ccc(O)cc3)c[nH]c2=O)C=C(C)C2CCC(C)CC21

CN(NC(=S)Nc1ccc(Oc2ccccc2)cc1)C(=S)Nc1ccc(Oc2ccccc2)cc1

CC1=C(c2ccc(C(C)C)cc2)S(=O)(=O)N=C1N=c1ccc(Br)c[n-]1

O=C1CC(c2ccc(Cl)cc2)CC2=C1C=Nc1ccccc1N2

CC=CC(C(=O)NC(C)(C)CC(C)(C)C)N(CCCCCC)C(=O)C1CC2CC(C1C)C2(C)C

Cc1c(C)c(C)c(C(=O)c2c(Cl)c(Cl)c(Cl)c(Cl)c2C(=O)[O-])c(C)c1C

CCC1(c2ccc(F)cc2)NC(=O)N(CC(O)COC(c2ccc(F)cc2)c2ccc(F)cc2)C1=O

CN(CCCF)c1ccc(-c2cn3cc(I)ccc3n2)cc1

Nc1ccc(Cl)c(Cn2nc([O-])ccc2=O)n1

O=C1N=C([O-])C2C1C(C(=O)OC(c1ccccc1)c1ccccc1)N1c3ccccc3C=CC21

FC(F)(F)c1ccn(Cc2nnc([S-])n2-c2cnn(Cc3ccc(Cl)cc3Cl)c2)n1

CCN1C(=CC=Cc2sc3cc(C=CC(=O)[O-])ccc3[n+]2CC)Sc2cc(C=CC(=O)[O-])ccc21

COc1ccccc1C1(C2=C(OC(C)C)C([O-])(c3ccccc3)C2=O)SCCCS1

CC(=CC(=O)CC(C)(O)C1C(O)CC2(C)C3CC=C4C(C=C(O)C(=O)C4(C)C)C3(C)C(=O)CC12C)CO

Cc1ccc(SC2=CC(=Nc3ccccc3)c3c([O-])c4ccccc4c4onc2c34)cc1

Cn1c(=O)c2c(=NOCCO)cc[nH]c2n(C)c1=O

CSC(=CC(=C(C(=O)[O-])C(=O)[O-])c1ccc(Br)cc1)SC

[O-]c1nc(N(CCO)CCO)nc2c(N3CCCCC3)nc(N(CCO)CCO)nc12

Nc1c(Nc2cc(F)c(F)cc2F)ccc2scnc12

CCOC(=O)c1cnc(SCC(=O)OCC(=O)N(C2CCCCC2)C2CCS(=O)(=O)C2)nc1N

CCc1ccc(NC2=C3CC(C(C)(C)C)CC=C3N(c3ccc(CC)cc3)C2=O)cc1

Cc1nc(-c2ccc(CNC(=O)C(C)SCc3nc4sc(C)c(C)c4c(=O)[nH]3)s2)cs1

Smiles String

CC1(C)C2CCC1(CS(=O)(=O)N1CCC3(CCc4ccccc43)CC1)C(NC(=O)C1CNC(=O)N1)C2

CN(C)c1ccc(C=Nn2c(N)c(S(=O)(=O)c3cccs3)c3nc4ccccc4nc32)cc1

CC([NH3+])C(Sc1nnc2ccccn12)c1ccccc1Br

C=C1C=C([O-])N2C(=NC(Nc3ccc(OC)cc3)=NC2c2cccc(Br)c2)N1

NC1=NC(=O)C(=NNc2ccc3c(c2)COC3=O)C([O-])=N1

CC1CCCC([NH3+])(c2noc(-c3c(Cl)cccc3Cl)n2)C1

Cc1ccc(C(C(C#N)c2ccc(Cl)cc2Cl)N2CCOCC2)cc1

Cc1ccccc(OC(C)C(=O)NNC(=O)CCCCCNC2=NS(=O)(=O)c3ccccc32)c1C

CCc1[nH]c(CN2Cc3ccccc3OC(c3ccccc3F)C2)c(C)[nH+]1

C=CC1=C(C(=O)[O-])N2C(=O)C(NC(=O)C(=NO)c3csc(N)n3)C2SC1

CC(Sc1nnc(-c2ccc(O)cc2)c(-c2ccc(O)cc2)n1)C(=O)c1ccc(N2CCCC2=O)cc1

CC1NNC(=O)C2=C1C(C)OC2NC(=O)CSC1NN(c2ccccc2)C(=S)S1

Cc1nccc(CNc2c(F)c(F)c(F)c(F)c2F)n1

CC(C)CN(CC(C)C)C(=O)C12CCC(C)(C(=O)C1Br)C2(C)C

O=C([O-])CCC=CCC1COC(c2ccccc2Cl)OC1c1cccnc1

O=CCC(NC=O)c1ccc(Cc2ccccc2)c(Cc2ccccc2)c1S(=O)(=O)Oc1ccccc1

C=C(NN1C(=Nc2cccnc2)SCC1(O)c1ccc2c(c1)NC(=O)CO2)c1ccc(C)c([N+](=O)[O-])c1

Cc1cc2c3c([nH]c2cc1S(=O)(=O)NCCC[NH+]1CCCC(C)C1)CC(C)(C)CC3=O

Nc1ccccc1SC(C=CNNC1=[NH+]C(c2ccccc2)C(c2ccccc2)=NN1)c1ccccc1

N#CC1=C([S-])[NH+]=C(N)C(c2nc(-c3ccccc3)cs2)C12CCCCC2

Fc1cnc(-c2cc(-c3cnnc(-c4c(F)cc(F)cc4F)c3)ccc2F)c(F)c1

C#CC[NH2+]C1CC(c2ccc(OC(F)(F)F)cc2)N2C(C(=O)OC)Cc3c([nH]c4ccccc34)C2C1

COCOCC(C)(O)c1ccccc1CCC(O)c1cccc(C=Cc2ccc3ccc(Cl)cc3n2)c1

C=Cc1c(C)c2cc3[nH+]c(cc4[nH]c(cc5[nH+]c(cc1[nH]2)c(C)c5CCC(=O)[O-])c(CCC(=O)[O-])c4C)c(C)c3C=C

Cc1cc(NC(=O)CC2CCCC2)n(-c2nc([O-])c3cnn(-c4ccc(Cl)cc4)c3n2)n1

CC(C)S(=O)(=O)C1CN(c2cc(-c3ccccc3)ccc2[N-]S(=O)(=O)S(=O)(=O)C(C)C)CCN1

CC(C)(CC#N)c1cc(CC(C)(C)c2cc(CC(C)(C)c3cc(CC(C)(C)C4(C)OCCO4)no3)no2)no1

CCN(CC)S(=O)(=O)c1ccc(Nc2ncnc(N(CCC#N)CCC#N)c2N)cc1

COc1c(CC2CC(c3ccc(Cl)cc3)=NO2)c(C=NNC(N)=S)c2c(c1OC)OCO2

[NH3+]C(CO)(CO)CCc1ccc(OCCCCCCNc2ccc([N+](=O)[O-])c3nonc23)cc1

CC1=CC(=N)SC1c1nncn1C1CCCCC1

Cc1ccc(NC2CC3=C(CC2(C)C)C(C)(C)CCC3)cc1

Cc1cc(OCC(=O)OC(C)(C)C)cc2c1C(=O)C(=Cc1cc(Br)cc3c1OCOC3)O2

Cc1ccc(C2(C(F)(F)F)NC(=O)C(C#N)=C3Sc4ccccc4N32)cc1

C=C(C(=O)[O-])C1CCC2(C)CCC(OC(=O)C(=C)C3CCC4(C)CCCC(C)(O)C4C3)OC(C)=C2C1

O=C([O-])CC(O)CC(O)C=Cc1c(C2CC2)nc2sccc2c1-c1ccc(F)cc1

CCC1C=C2C=C(C(=O)NCCC3=c4cc(C)ccc4=[NH+]C3C)SC2S1

OC(c1cnc(NC(c2ccccc2)(c2ccccc2)c2ccccc2)s1)(C(F)(F)F)C(F)(F)F

CC(Oc1cccc(I)c1)(C(=O)[O-])c1c(F)cccc1F

CC1=CC2C=C(C)C(C)C3C(CC(C)C)NC(=O)C23OC(=O)C=CC(O)CCC1

O=NN(CCF)C(=O)NC1CCC(=O)NC1=O

Smiles String

Cc1cn(CCCCCCCCCCCCOc2c3occc3cc3ccc(=O)oc23)c(=O)[nH]c1=O

Cc1nnnn1N=Cc1ccc(O)c(C(c2ccc(O)cc2)c2cc(C=Nn3nnnc3C)ccc2O)c1

Oc1ccc(Br)c(C2Nc3ccccc3-c3nnc(SCc4ccccc4F)nc3O2)c1

CCS(=O)(=O)c1nnn(-c2nnc(C)s2)c1N

S=c1oc(CSc2nnc(-c3ccncc3)n2-c2ccccc2)nn1C[NH2+]CCN1CCOCC1

COC(=O)C(CC(C)C)NC(C)c1ccc(Cl)c(Cl)c1Cl

COc1cc(OC)c(N=C(Nc2cc(Cl)c(OC)cc2OC)SC2C=CCCC2)cc1Cl

CCc1ccccc1-n1nc(C(=O)N(C)CCc2c[nH]c3ccccc23)cc1-c1cccc(NC(C)=O)c1

Nc1c[nH+]ccc1N1CCN(Cc2ccc(Br)cc2)CC1

CCC=CCC1=C(C)C(n2ccc3ccccc32)(n2ccc3ccccc32)CC1

COc1cc([N+](=O)[O-])cc(C=Nc2ccc3oc(-c4ccc(Br)cc4)nc3c2)c1[O-]

CCC(C)=NCCO

O=C(NCC1(O)CCCCC1)c1ccc(C2CCC[NH+]2C2CCC2)s1

NC(=O)C1CN(C(=O)COC(=O)c2ccccc2OCc2ccc(Cl)cc2)c2ccccc2O1

COc1cc(C(c2c([O-])cc(=O)n(C)c2C)c2c([O-])cc(C)n(C)c2=O)cc(OC)c1OC

CCOC(=O)C1N=C(C)C(CCC(=O)N2CCC(C(=O)NCCCN3CC[NH+](C)CC3)CC2)=C1C

CCCCCc1ccc(SCc2nnc(CSc3ccc(CCCCC)cc3)n2N)cc1

Cc1ccc(C)c(N(C(=O)c2snc(C(N)=O)c2N)C(C(=O)NC2CCCC2)c2ccc(C)o2)c1

O=c1cccc2n1CC1CC2CN(CC2C[NH+]3CCC2CC3CNC(=S)NCc2ccco2)C1

Cc1cc(C(=O)CSc2nnc(-c3cccs3)n2CC(=O)[O-])c(C)n1CCC1=CCCCC1

CCCNC(c1ccc(Br)s1)c1nc(CC)c(C)s1

COC(=O)C(Cc1c[nH]cn1)NC(=O)CONC1=CCC2(C)C(=C1)CCC1C2C(O)CC2(C)C1CCC2(O)C(=O)CO

Nc1ccc(Oc2ccc(Cl)c(Cl)c2)c2ccncc12

CCN1c2c(OCCC[NH2+]CCc3c[nH]cn3)c[nH+]c(-c3cccc(Cl)c3)c2NC1c1nonc1N

Cc1ccc2c(Nc3ncnc(NNC(=O)C(c4ccccc4)c4ccccc4)c3[N+](=O)[O-])cccc2[nH+]1

O=C(CSC1CCOCC1)c1cc(Br)sc1Br

Cc1nn(-c2ccccc2)c2c1C(c1ccc(C(C)C)cc1)N1C(=N2)C([N-]c2cc(Cl)cc(Cl)c2)=Nc2ccccc21

CC(C)=CCCC(=O)C(c1nc2nnc(N)n2nc1[O-])c1nc2nnc(N)n2nc1[O-]

C[NH+]1CCC23C=CC(O)CC2Oc2c(OC4OC(C(=O)[O-])C(O)C(O)C4O)ccc(c23)C1

CCc1cccc(CC)c1NC(=O)c1c(NCc2ccco2)sc(C(=O)Nc2cc([N+](=O)[O-])ccc2Cl)c1N

O=C(C1=NNCC1c1cc2ccccc2c2ccccc12)c1cc2ccccc2c2ccccc12

CC=C(C)C=CC=CC(OC)C(C)C(OC)C(C)CCc1oc2c(O)c(OC)cc(OC)c2c(=O)c1C

CC(C)S(=O)(=O)N1CC(C)(S(=O)(=O)C(C)C)N(CC2(C3CCCCC3)CCN(Cl)CC2)C1=O

Cc1ccsc1C1SCC(=O)N(CC(=O)NCC[NH+](C)C)c2c1c(-c1ccccc1)nn2C(C)(C)C

CCOc1ccc2c(c1)N(Cc1ccc(CN3C4=C(CCC(OCC)C4)C(C)=CC3(C)C)cc1)C(C)(C)C=C2C

Cc1ccc(-c2nnc(NNC3=CC([O-])(c4ccccc4)C(c4ccccc4)=C3)c3ccccc23)cc1C

CCC(C)C(C)n1c(=S)[nH]c2sccc2c1=O

CNc1ccc2oc(CC3OC4(CCC3C)OC(C(C)C(=O)c3ccc[nH]3)C(C)CC4C)nc2c1C(=O)[O-]

Cc1cc(N2CCC(CO)CC2)n2ncnc2c1

N#CC(c1ccc(C(C#N)=[N+]([O-])O)cc1)=[N+]([O-])O

CCCC1CC(C)(c2csc(NCCNc3nc(C4(C)CC(CCC)C(=O)O4)cs3)n2)OC1=O

Smiles String

Cc1nn(C(C)C)c(C)c1C=NNc1nc(SCC(=O)Nc2nnc(SCC(N)=O)s2)n[nH]1

[O-]c1nc(C(Cl)=Cc2csnn2)nc2scc(-c3ccccc3Cl)c12

C=CCOC(=O)C(CSC(=O)C1(C(O)C(C)C)NC(=O)C(C)C1O)NC(C)=O

CC1(C)COCCN1C(=O)NC(=O)CCCl

C[NH+]1CC2CC1CN2c1ccc(-c2ccc3[nH]c(C(F)(F)F)cc3c2)cn1

C=CC1(C)CC(OC(=O)CSCC[NH+](CC)CC)C2(C)C(C)CCC3(CCC(=O)C32)C(C)C1O

O=C(C=C(C1CC1)C1CC1)Nc1ccccc1I

CC1CCCC(N(C)C(=O)C(=O)NCC(N)=S)C1

CC[NH+](CC)c1ccc(C2C3=C(SC(=O)c4ccccc42)N(C(=N)Nc2ccccc2)CC(C)(C)C3)cc1

CC(C)(C)CC(C)(C)SSc1nnc([S-])s1

CC(=NN=c1[nH]c(-c2ccc(C)cc2)cs1)c1ccc(Cl)c(Cl)c1

CC(Sc1nnc(Cc2cccs2)o1)c1ccc(F)c(Cl)c1

OCC(CO)(CO)[NH2+]CC(O)Cn1c(-c2ccccc2)c(-c2ccccc2)c2ccc3ccccc3c21

Ic1ccc(C=Cc2ccc(-c3ncc(-c4ccccc4)o3)cc2)o1

CC(C)(C)C(=O)OCC1OC(F)C(OC(=O)C(C)(C)C)C(OC(=O)C(C)(C)C)C1OC(=O)C(C)(C)C

CCCn1nc2ccccc2c1P(=Nc1ccc(S(N)(=O)=O)cc1)(N1CCCCC1)N1CCCCC1

Cc1ccc(-c2nc([S-])nn2-c2nc3ccccc3s2)cc1

[O-]c1c2c(nc3cc(-c4ccco4)nn13)CSC2

C[NH+](C)CCCNC(=S)N(CCCO)Cc1cc2cc3c(cc2[nH]c1=O)OCO3

CN(Cc1csc(Br)c1)C(=O)C=Cc1ccccc1Br

CC1(C)CCC2(C(=O)[O-])CCC3=C(C(CCl)CC4C3(C)CCC3C4(C)CC(O)C(O)C3(C)C(=O)[O-])C2C1

CCSC1=NC(=O)C2C(=N1)N=C1CCCC([O-])=C1C2c1ccc(N(CC)CC)cc1

C1=C(c2ccccc2)N=NC1C1NNC2SC(COc3ccc(OCc4ccccc4)cc3)=NN21

Nn1c(Nc2ccc(S(=O)(=O)NC3C=CNO3)cc2)nnc(C=Cc2c(O)ccc3ccccc23)c1=O

CC(=Cc1ccccc1)c1nc2ccc(Br)cn2c1NC(C)(C)C

Cc1cc(I)ccc1NC(c1ccco1)c1[nH+]ccn1C

CC1=C(C(c2ccc(OC(=O)c3ccc(C(C)(C)C)cc3)cc2)c2c(C)n[nH]c2[O-])C(O)N=N1

CCC1c2nc(C(F)(F)F)n(C)c2CCN1C(=O)CC([NH3+])Cc1cc(F)c(F)cc1F

O=C(NCc1ccc(Oc2ccc(F)cc2)cc1)c1oc2ccccc2c1CSc1ncn[nH]1

Cc1c(CN(C)CC(O)c2ccc(O)cc2)sc2c(=O)c(C(=O)NCc3ccc(Cl)cc3)cn(C)c12

OC(C[NH2+]CC(O)(c1ccc(Cl)cc1)c1ccc(Cl)cc1)(c1ccc(Cl)cc1)c1ccc(Cl)cc1

Cc1c(C)n(C)c2ccc(C(=O)NC3C4CC5CC(C4)CC3C5)cc12

CC(C)(C)NC(=O)C1=C(Cl)C2C=CC=C(NC(=O)C(C)(C)Oc3ccc(Br)cc3)C2S1

C=CCC1C=C(C2=C(c3ccncc3)C(c3ccc(F)cc3)=NC2)CCN1C

C[NH2+]C1CCc2cc(OC)c(OC)c(OC)c2-c2ccc(NCCCN3CCCc4ccccc43)c(=O)cc21

Cc1cc(C)c(-c2c(Nc3ccc(F)cc3C)c(=O)c2=O)cc1C

O=P1(c2ccccc2)N=C(c2ccccc2)N=C(C(Cl)(Cl)Cl)[N-]1

CC1(NC(=O)Nc2ccc(O)c(NC(=O)NC3(C)CCS(=O)(=O)C3)c2)CCS(=O)(=O)C1

Nc1nc2c(nc(Sc3ccc(Cl)cc3)n2C2OC3COP(O)(=S)OC3C2O)c(=O)[nH]1

O=c1cccn[nH]1

CC(N)C#N

Smiles String

COC1CC(OC(C(C)C(=O)OC(C)C(C)C(O)C(C)C(=O)C2(C)CO2)C(C)C(OC2OC(C)CC([NH+](C)C)C2O)C(C)C)OC(C)C1O

O=[N+]([O-])c1ccc(NN=Cc2ccc(-n3ccnc3)cc2)c(S(=O)(=O)[N-]c2ccc(Cl)cc2)c1

CCCCN1c2ccc([N+](=O)[O-])cc2CC2(C(=O)N(c3ccc(F)cc3)C(=O)N=C2[O-])C1CCC

CC1=NC(C)=C(C(=O)OCC(C)C)C(C2=CC=CN(C(=O)OC(C)(C)C)C2)C1=C([O-])OCC(C)C

CC(O)c1ccc(NC(=O)C2=C(Br)C3=NC(c4cccs4)CC(C(F)(F)F)N3N2)cc1

COc1ccc(CCn2c(-c3ccc(Cl)s3)csc2=Nc2ccc(C)cc2C)cc1OC

CC(O)C1C(=O)N2C(C(=O)[O-])=C(SC3C[NH2+]C(CSc4nnnn4-c4ccncc4)C3)C(C)C12

CCCCCCSc1ncnc2c1sc1nc(CC(C)C)c3c(c12)CC(C)(C)OC3

CCCN(C(=O)c1cc2cc3ccc(OC)cc3nc2s1)c1ccc(CC)cc1

COc1cc2c(C#N)c(N(C)C(c3ccccc3)c3ccccc3)nc(N)c2c(N)n1

Fc1ccc(C2CC(c3cccs3)=NN2c2nc(-c3ccc(Br)s3)cs2)cc1

Cc1ccc(C2CCC[NH+]2CN2N=C(c3ccc(F)c(C)c3)N[N-]2)cc1

O=S(=O)(NC(c1ccccc1)C(c1ccccc1)n1nnc2ccccc21)c1ccccc1

CC1CN(c2ccc(N=C([S-])Nc3ccc(N4CC(C)OC(C)C4)cc3)cc2)CC(C)O1

CC1=C(C(=O)c2ccccc2)C(c2ccccc2)C(C(=O)c2ccccc2)C(C)N1

Cc1cc(C#N)c([S-])nc1-c1c(C)c(C)cc(C)c1C

CCCCn1c(=O)[nH]c(=O)c2c1nc(CCC(=O)NNC(=O)C(=O)NC(C)C)n2CC(C)C

CCCC12CN3CCC(CC)(C[NH+](C1)C3)C1C(=O)Nc3ccc(Br)cc31)C2O

CC(=NNC(=O)c1sc2ccccc2c1Cl)c1ccc(NC(=O)C2CC=CCC2C(=O)[O-])cc1

Clc1ccccc1C(C[NH+]=Cc1ccncc1)c1c[nH]c2ccccc12

CN(C)c1cc(NN=Cc2cn(CCOc3ccc(C4CCCCC4)cc3)c3ccccc23)[nH+]c(N(C)C)n1

Cc1cc(C(C)(C)C)[n+](-c2ccc(S(=O)(=O)Nc3nnc(S(N)(=O)=O)s3)cc2)c(C(C)(C)C)c1

O=S(=O)([N-]c1cc(-c2nnc3n2CCCCC3)c(F)cc1F)c1ccc(F)c(F)c1F

Cc1ccccc1-c1c(-c2cc3c(s2)CCC3)n[nH]c1N

Cc1ccc(C)n1-c1ccc(Cl)c(C(=O)NN=Cc2cc(Br)c(N3CCCC3)o2)c1

N#CC1=C(SCc2nc3ccccc3[nH]2)NSC1SCc1nc2ccccc2[nH]1

COc1cccc(C=NC23CCCCC2C([NH+]=Cc2cccc(OC)c2O)c2cccc(OC)c2O3)c1O

C=C1C=C(c2ccc(Br)cc2)C(C#N)C([S-])=N1

CC(O)C(O)C=CC=CC(=O)OC1C=C2COC(O)C2(O)C2(C)CCCC(C)(C)C12

COc1ccc(C(C)=NOCc2nc3c4c(-c5ccc(OC)cc5)c(-c5ccc(OC)cc5)oc4ncn3n2)cc1

CCOCC(C)[NH2+]C1C(=O)Nc2c(C)cc(Br)cc21

S=C(NC1CC1)c1c(-c2ccccc2)c2c3n(c(-c4ccc(Cl)cc4)cn13)CCCC2

CC1(C(=O)NN)CC=CC=C1C(=S)OC(=O)C=C([O-])OC(=S)C1=CC=CCC1(C)C(=O)NN

CCOC(=O)C(C#N)=CC1=NC(C(c2ccc(N(C)C)cc2)c2ccc(C=C(C#N)C(=O)OCC)[nH]2)C=C1

O=C([O-])c1c[nH]nc1S(=O)(=O)NCc1nc[nH]n1

N#Cc1cccc(C[NH2+]C2CCC(Nc3ncc(-c4ccsc4)c4nc(C(F)(F)F)ccc34)CC2)c1

N#CC(=Cc1ccc(OCc2c(Cl)cccc2Cl)cc1)c1nn(CCO)c(N)c1C#N

Cc1cc(C)cc(NC(=S)NC(=Nc2nc(C)cc(C)n2)N2CCC(C(N)=O)([NH+]3CCCCC3)CC2)c1

OCCn1nccn1

C#CCC(O)CC

Smiles String

CCI=NC=CCI

CC(C)CCINc2cc(NS(=O)(=O)N(C)C)cc(C(=O)NCC3C=c4ccccc4=[NH+]3)c2NIC

CcInc(-c2ccccc2)nc(N2CCN(CC(O)COCC(C)C)CC2)cICcIccc([N+](=O)[O-])ccI

CCC(OC)C([O-])=NCI=NC2=C(CCCC2)C2(CCCCC2)SI

CCCN(CC(=O)NcIccc(F)c(F)cIF)C(=O)CScIccc(C(N)=O)ccI[N+](=O)[O-]

CCICC(OCC(O)C[NH+]2C(C)(C)CC(O)CC2(C)C)CC(C)(C)CI

NC(=O)NICCCC(C(=O)N2CCCCC2CO)CI

CC[NH+](CC)cIccc(-n2nc3cc(C)c(NC(=S)NC(=O)c4ccc(OC(C)C)cc4)cc3n2)c(C)cI

NcIc2c(-c3ccc(F)cc3)nc(N3CCN(CC[NH+]4CCCCC4)CC3)nc2nnI-cIcccc(F)cI

C=CCN(CC(=O)NICCc2sccc2CIcIccc(C(C)(C)C)cc1)C(=O)cIccc(C(C)(C)C)ccI

CCCN=C(CCC)CI=C([O-])CC(CC(C)SCC)C(C(=O)OCC)CI=O

C=C(C)CIC(O)CC2(C)Oc3c(C)c4c(c(OC)c3CC2CI(C)CCC(=O)[O-])C(=O)OC4

N=CISCC(=O)NIcIncc(Cc2ccc(Cl)c(Cl)c2)sI

CCICCICI(O)CCI

C=CCnIc(Cc2csc(NC(=O)CCC3[NH+]=C(c4ccc(OC)cc4)NO3)n2)n[nH]cI=S

NC(=NcIccc(Cl)ccICl)N=cInc([O-])cc(CSc2ccccn2)[nH]I

CC(C)=CCIC(C(=O)SCc2cc3c(cc2Cl)OCO3)CI(C)C

CCCICCC(NC(=O)C2CCC[NH2+]2)(C(=O)[O-])CCI

O=P([O-])(OC(cIcccccI)C(F)(F)C(F)(F)C(F)(F)F)cIccccI

NNcInc2sccn2cIS(=O)(=O)NcIccc(I)ccI

CCScIcccccIC(=O)NICc2nn(-c3ccc(C)c(C)c3)c(-n3cccc3)c2CI

O=[N+]([O-])C(Cl)=CC(=CINCCCNI)[N+](=O)[O-]

FcIccc(C2=NN=CC2=CNN=c2cc[nH]c3cc(Cl)ccc23)ccI

CCOC(=O)CNIC(=O)C2(C(C(=O)OCC)=C(N)OC(CC(=O)OC)=C2C(=O)OC)c2ccccc2I

C(=NcIccccI)CI=C(c2ccccc2)C(=CNc2ccccc2)CCCI

CCCCOC(=O)CNC(=O)OCC=CC#CC#CC=CCOC(=O)NCC(=O)OCCCC

CCC(C)(C)cIccc(OCCCCNC(=O)c2coc(C3CCC[NH2+]C3)n2)c(C(C)(C)CC)cI

O=CICCCc2nc(NC(=O)C3CCCC3)sc2I

CCI=NC2C=CC(C3NNC(SCC(N)=O)N3c3ccc(Br)cc3)=CC2=CIC

CC(C)(C)cIcc(=O)[nH]c(C23CC4CC(CC(C4)C2)C3)nI

CCOcIccc(OCC=C(C)CCC=C(C)CC)ccI

CC(C=NNC(=O)CNcIccc(I)c(C)cI)=CcIccc(C(C)C)ccI

CCCc1nc(N2CCN(C([O-])=NS(=O)(=O)c3ccc(C)cc3)CC2)c2c3c(sc2n1)CC(C)CC3

C[NH+](CC(=O)NCCSCcIccccIC#N)CcIccc(C(=O)NCC(F)(F)F)ccI

CcIcc(C(=O)C(C#N)=C([O-])c2sccc2C)scIC

O=C(N[NH+]=CcIccccI[O-])cIc([O-])nnc(-c2ccccc2)cI-cIccccI

CCCCCI=C(Cc2ccc(-c3ccccc3-c3nnn[n-]3)cc2)C(C(=O)[O-])N(c2ccccn2)NI

NcIccc(SC2CC([O-])=NC2=O)ncI

CNC(=O)CnIcc(NC(=O)c2csnn2)cnI

CCICC(=Cc2ccc(Sc3nncn3C)c([N+](=O)[O-])c2)c2nc3ccccc3c(C(=O)[O-])c2CI

COC(=O)C(CCCCnIc(-c2csc(COC(=O)C(C)(C)C)n2)cc(C(N)=O)cIC)NC(C)=O

CC(NC(=O)CI(O)CCC(F)(F)CCI)cIncc(-c2cc(Cl)cc(F)c2-c2nnn(C)n2)ccIF

Smiles String

O=S1(=O)CCC([NH+](Cc2ccc(-c3cccc(Cl)c3)o2)CC(O)COc2ccccc2F)C1

CCOCCCN(CC(=O)N(Cc1ccccc1)Cc1ccc(C)o1)S(=O)(=O)C=Cc1ccccc1

O=C([O-])C(CNC(=O)C1CCCn2c(C=CC3CC[NH2+]CC3)nnc21)NS(=O)(=O)Cc1ccccc1

CCOC(=O)c1c(-c2ccc(C)cc2C)csc1NC(=O)COc1ccc(C(C)(C)CC(C)(C)C)cc1

CCCc1(C)C(NC(C(Oc2ccc(Cl)cc2)n2cncn2)C(C)(C)C)=[NH+]C(=S)N1c1ccc(C)cc1

N=C1N=C([O-])C2=NCN(C3OC(CO)C(O)C3O)C2=N1

NS(=O)(=O)c1ccc([NH+]=c2oc3cc(O)ccc3cc2C(=O)Nc2ccc(Cl)cc2)cc1

Cc1c(C(=O)NNC(=O)c2cc(=O)[nH]c3ccccc23)sc2c1C(=O)NNS2(=O)=O

CC(C)(C)c1cc(C(=S)N2CCCCCC2)cc(C(C)(C)C)c1O

CC1(Cc2ccc(Br)cc2)C(=O)OC(c2ccc(Br)cc2)C(C)(C)C1=O

OCCCNc1ncnc2c1[nH]c1ccc(Br)cc12

CCn1c([S-])nnc1COc1ccc(Br)cc1C(C)C

O=C1NC(=O)N(c2ccc(Br)cc2)C(=O)C1C(=NCCc1c[nH]c2ccc(F)cc12)c1ccc(Cl)cc1

COc1ccc(C2(c3cccc(-c4cccnc4F)c3)N=C(N)N3CCC[NH+]=C32)cc1

NC(=O)C[NH+]1CCCN(c2ncccc2CNc2nncn2-c2cccc(Cl)c2Cl)CC1

Cc1cc(C)c(Cn2cccc2C2=NC(CSCc3ccc(F)cc3)CO2)c(C)c1

Cc1cc(S(=O)(=O)Oc2ccc(C3SCCS3)cc2)c(C)s1

O=C(Nc1nc(-c2cccc(C(F)(F)F)c2F)cs1)c1ccc(Nc2cccnc2)cc1

CSc1nccc(-c2cc(Cl)c3c(c2)CC(CNC(=O)CC2CC(C)NC(=S)N2)O3)n1

Cc1ccc2c(c1)C(O)CC1(CC[NH+](CC(O)c3ccc4c(c3)CCC4)CC1)O2

CC1CC(C)CN(S(=O)(=O)N2CCCC(O)C2)C1

COC(=O)C1Sc2nc([S-])sc2C(c2ccc(Cl)cc2)C1C(=O)OC

O=C1C(=CC=Cc2ccccc2)[N-]C(c2ccccc2)N1c1c(Cl)cc([N+](=O)[O-])cc1Cl

CCOC(=O)C1=C([O-])C(=Cc2cc(Cl)c(OCC#N)c(Cl)c2)SC1=Nc1ccc(CC)cc1

O=C1CC(NN2C(=O)CC(c3ccc(C(F)(F)F)cc3)C3=C2CCCC3=O)=C(c2ccccc2)N=N1

C=C1C(=O)C23C4CC5C(C)(C)CCC(OC(C)=O)C5(C)C2C(OC(C)=O)CC1(OC(C)=O)C3OC(C)(C)O4

O=C([O-])C1(Oc2c(Br)cccc2Br)CCCSC1

Cc1c(Cl)c(C)c([N+](=O)[O-])c([O-])c1C=Nc1ccc2c3c(cccc13)CC2

O=C([O-])CC12CC3CC(C1)CC(n1ncc(Nc4cccc(C(F)(F)F)c4)c(Cl)c1=O)(C3)C2

CC1CCCC(Nc2c(-c3ccc(F)cc3)[nH]c3c(C#N)cnn23)C1C

CCC[NH+]1CCC=C(c2ccc(Nc3nc(Nc4ccccc4C(N)=O)c4cc[nH]c4n3)c(C)c2)C1

COC(=O)C=c1sc2n(c1=O)C(N)=C(C#N)C(c1ccc(F)cc1)C=2C(=O)Nc1ccc(C)c(C)c1

Cc1cc(Oc2cccc(C(C)C)c2)nc(SCc2cc(F)cc(F)c2)n1

CCc1cccc(CC)c1N1C(=O)C2C(c3ccc(OC(C)=O)cc3)[NH2+]C(C(=O)[O-])(c3ccccc3)C2C1=O

Cc1ccccc1NC(=S)NC(C)C(c1cccs1)N1CCN(c2ccc(F)cc2)CC1

C=Cc1ccc(COc2ccc(-c3n[nH]c(C(F)(F)F)c3-c3ccc(OC)c(OC)c3)c(O)c2)cc1

CCOc1cc(CCc2nc(C)c(CC)s2)[nH+]c(NCc2cc(Cl)cc(NC(=O)OC(C)C)c2)c1

CC1CCC(C(N)=O)C[NH+]1CC=CC(=O)[O-]

COc1ccccc1C=CC[NH+]1CCNC(=O)C1CC(=O)NCc1cc2cc(F)ccc2[nH]1

CC1=C(C(c2ccccc2O)c2c(C)[nH]n(-c3ccccc3)c2=O)C(O)N(c2ccccc2)N1

CC(C)(Cn1c([S-])nc2ccccc2c1=O)[NH+]1CCc2ccccc2C1

| Smiles String |
| --- |
| CN(C)S(=O)(=O)N=C([O-])C=Cc1ccc(SC(F)(F)F)cc1 |
| CCOc1cc(-c2ccccc2O)cc2c1N(C)c1ncnc(Cl)c1N=C(c1ccccc1O)C2 |
| Cc1sc2nc(C[NH+]3CCCCC3)nc(NCCC(O)(c3nccn3C)C(F)(F)F)c2c1C |
| O=C(c1ccc(Cl)cc1)C1C(c2ccc(-c3ccccc3[N+](=O)[O-])o2)NC(O)=[NH+]C1([O-])C(F)(F)F |
| CC1(C)CC(NC(=O)N2CCOC(c3ccc(F)c(F)c3)C2)CC(C)(C)[NH2+]1 |
| COc1ccc(C(=O)C=Cc2cn(CCC#N)nc2-c2ccc(F)c(F)c2)cc1CSCC(=O)[O-] |
| COC(=O)C1=C(C)N=C2SC=C(C=C([O-])NCc3ccncc3)N2C1c1ccsc1 |
| NC(=O)COc1ccc(C2c3c(oc4ccc(Br)cc4c3=O)C(=O)N2Cc2ccco2)cc1 |
| Cc1[nH]c2c(C(=O)N3CCN(CCOc4cccc(C[NH+](C)CC5CCC(=O)N5)c4)CC3)cccc2c1C |
| |
| C=C(C)C1CCC2(C(=O)OC)CCC3(C)C(CCC4C5(C)CC(=C[O-])C(=O)C(C)(C)C5CCC43C)C12 |
| C=CCc1ccc(O)c(-c2cc(C3Oc4c(-c5cc(CC=C)ccc5O)cc(O)cc4C3CO)ccc2O)c1 |
| COc1ccccc1C1C2=C(N=c3sc(=Cc4ccc(Sc5ccc(Cl)cc5)o4)c(=O)n3I)c1ccccc1CC2 |
| Cc1ncsc1C(=O)N(C)CC(N)=NO |
| COc1ccc(-n2c(NN=CC3=Cc4ccccc4OC3C)nc3c(c2=O)C(C)(C)Cc2ccccc2-3)cc1 |
| O=C1C2CC=C(Cl)CC2C(=O)N1c1cc[nH]n1 |
| CCCCCC12CCC(C(=O)Oc3ccc(N=C=S)cc3)(CC1)CC2 |
| CC(C)=CC(=O)NCCCC(NC(=O)C1C(NC(=O)NC(C)C)CCN1C(=O)C=C(C)C)C(N)=O |
| CC(Nc1nccc(-c2c(-c3ccc(F)cc3)nc3cc(C[NH+](C)C(CO)CO)ccn23)n1)c1ccccc1 |
| COc1cc(C=NNC(=O)c2nnn(-c3nonc3N)c2C[NH+]2CCCCC2C)cc(OC)c1OC(C)=O |
| COC(=O)C(O)C(O)(CCC(C)C)C(=O)OC1C(OC)=CC23CCCN2CCc2cc4c(cc2C13)OCO4 |
| COc1ccc2c(C)cc3nnc(SCC(O)COc4cc(C)ccc4C)n3c2c1 |
| CC(C)(F)CC(NC(c1ccc(-c2ccc(C3(O)CC3)cn2)cc1)C(F)(F)F)C(=O)NC1(C#N)CC1 |
| C#Cc1ncnn1C |
| c1ncc2c(n1)NCC2 |
| C=CCOc1cccc(NC(=O)C2CCC[NH+](Cc3nc(-c4cccc(Br)c4)no3)C2)c1 |
| O=C1c2c(O)cccc2-c2nn(CC[NH2+]CCO)c3ccc(NCC[NH2+]CCO)c1c23 |

Table B.2.3 SMILES of decoys used as candidate compounds in Data Set 1.

| PDB Code | Ligand Residue Code | Resolution / Å |
|---|---|---|
| 1d4y | TPV | 2.0 |
| 1hp0 | AD3 | 2.1 |
| 1ztz | CB5 | 2.2 |
| 2pk5 | 75 | 1.9 |
| 2pk6 | O33 | 1.5 |
| 3i8w | CB5 | 1.7 |
| 3kdc | JZP | 2.2 |
| 3kdd | JZQ | 1.8 |
| 3th9 | 9Y9 | 1.3 |

Table B.2.4 PDB structures used to generate CRANkS grids for Data Set 2.

| Chain | PDB Identifier | SMILES of Ligand |
|---|---|---|
| A | 5pnx | Cc1cc(N)no1 |
| A | 5poo | Cc1cc(NC(C2CC2)=O)on1 |
| B | 5po1 | Cc1cc(NCc2cccs2)no1 |
| A | 5po2 | C1C(Nc2ccc(cc12)O)=O |
| B | 5po3 | Cc1cc(NC(c2cccc2)=O)on1 |
| A | 5po6 | c1cc(ccc1c1cn[nH]c1N)[Br] |
| A | 5po7 | CNc1cc2c(cc1[N+](O)=O)N(C)C(N2C)=O |
| B | 5po7 | CNc1cc2c(cc1[N+](O)=O)N(C)C(N2C)=O |
| A | 5po8 | CN1C(CCc2cc(ccc12)N)=O |
| B | 5po8 | CN1C(CCc2cc(ccc12)N)=O |
| A | 5po9 | Cn1c(c(c2ccccc2)c(C(F)(F)F)n1)N |
| B | 5po9 | Cn1c(c(c2ccccc2)c(C(F)(F)F)n1)N |
| A | 5pob | CN1C(N(C)c2cc(ccc12)N)=O |
| B | 5pob | CN1C(N(C)c2cc(ccc12)N)=O |
| A | 5poc | CN1C(C=Cc2cc(c(cc12)[Br])[N+](O)=O)=O |
| A | 5pod | CN1C(C=Cc2cc(ccc12)N)=O |
| B | 5pod | CN1C(C=Cc2cc(ccc12)N)=O |
| A | 5poe | CN1C(C=Cc2cc(ccc12)N)=O |
| B | 5poe | CN1C(C=Cc2cc(ccc12)N)=O |
| A | 5pof | CC(NC1CCN(CC1)Cc1ccccc1)=O |
| B | 5pof | CC(NC1CCN(CC1)Cc1ccccc1)=O |
| A | 5pog | CNC(c1cc2cccnc2s1)=O |
| B | 5pog | CNC(c1cc2cccnc2s1)=O |
| A | 5poh | CC(NCCNc1ccccc1)=O |
| B | 5poh | CC(NCCNc1ccccc1)=O |
| A | 5poi | CNC(c1ncccn1)=O |

| Chain | PDB Identifier | SMILES of Ligand |
|---|---|---|
| B | 5poi | CNC(c1ncccn1)=O |
| A | 5poj | CC(NC1CCN(CC1)Cc1ccccc1)=O |
| B | 5poj | CC(NC1CCN(CC1)Cc1ccccc1)=O |
| A | 5pok | CC(N1CCN(CC1)c1ccccc1)=O |
| B | 5pok | CC(N1CCN(CC1)c1ccccc1)=O |
| A | 5pol | CC(N1CCc2cc(c(cc2C1)OC)OC)=O |
| B | 5pol | CC(N1CCc2cc(c(cc2C1)OC)OC)=O |
| A | 5pom | CNC(c1cc2ccccc2nc1)=O |
| B | 5pom | CNC(c1cc2ccccc2nc1)=O |
| A | 5pon | CC(N1CCN(CC1)c1ccccn1)=O |
| B | 5pon | CC(N1CCN(CC1)c1ccccn1)=O |
| A | 5poo | CC(N1CCN(CC1)c1ccc(C#N)cn1)=O |
| B | 5poo | CC(N1CCN(CC1)c1ccc(C#N)cn1)=O |
| A | 5pop | CC(N1CCN(CC1)c1ccncc1)=O |
| B | 5pop | CC(N1CCN(CC1)c1ccncc1)=O |
| A | 5poq | CC(N1CCN(CC1)[S+](C)([O-])=O)=O |
| B | 5poq | CC(N1CCN(CC1)[S+](C)([O-])=O)=O |
| A | 5por | CC(N1CCNC(C1)=O)=O |
| B | 5por | CC(N1CCNC(C1)=O)=O |
| A | 5pos | CC(NCc1ccc(cc1)OC)=O |
| B | 5pos | CC(NCc1ccc(cc1)OC)=O |
| A | 5pot | CC(NCc1cccc(c1)[Cl])=O |
| B | 5pou | CNC(NC1CCN(CC1)Cc1ccccc1)=O |
| A | 5pov | CNC(c1cc2cccnc2s1)=O |
| B | 5pov | CNC(c1cc2cccnc2s1)=O |
| A | 5pow | CC(NCc1ccccn1)=O |
| B | 5pow | CC(NCc1ccccn1)=O |
| A | 5pox | CCOC(C1CCc2c(C1)cn[nH]2)=O |
| A | 5poy | CC(NCCNc1ccccc1)=O |
| B | 5poy | CC(NCCNc1ccccc1)=O |
| A | 5poz | C[S+](NCC1CC(N(C1)C1CCCC1)=O)([O-])=O |
| B | 5ppo | CNC(c1cccnc1N)=O |

Table B.2.5 PDB structures used in the construction of CRANkS grids for BRD1.

# Chapter 3

| UniProt Code | PDB Code | Organism | Reason for Rejection |
|---|---|---|---|
| P07550 | 3ny8 | HUMAN | N/A |
| P07550 | 3d4s | HUMAN | N/A |
| P07550 | 3ny9 | HUMAN | N/A |
| P07550 | 3pds | HUMAN | N/A |
| P07550 | 3nya | HUMAN | N/A |
| P07550 | 2rh1 | HUMAN | N/A |
| P07550 | 3pog | HUMAN | N/A |

Table B.3.1 PDB structures for use in the ADRB2 dataset.

| UniProt Code | PDB Code | Organism | Reason for Rejection |
|---|---|---|---|
| P10275 | 1e3g | HUMAN | N/A |
| P10275 | 1gs4 | HUMAN | N/A |
| P15207 | 1i37 | RAT | N/A |
| P15207 | 1i38 | RAT | Identical ligand to 1i37 |
| P10275 | 1t5z | HUMAN | Identical ligand to 1i37 |
| P10275 | 1t63 | HUMAN | Identical ligand to 1i37 |
| P10275 | 1t65 | HUMAN | Identical ligand to 1i37 |
| P10275 | 1xj7 | HUMAN | Identical ligand to 1i37 |
| P15207 | 1xnn | RAT | N/A |
| P10275 | 1xow | HUMAN | Identical ligand to 1e3g |
| P10275 | 1xq3 | HUMAN | Identical ligand to 1xq3 |
| P10275 | 1z95 | HUMAN | N/A |
| P10275 | 2am9 | HUMAN | Identical ligand to 1i37 |
| P10275 | 2ama | HUMAN | Identical ligand to 1i37 |
| P10275 | 2amb | HUMAN | N/A |
| P10275 | 2ao6 | HUMAN | N/A |
| P10275 | 2ax6 | HUMAN | N/A |
| P10275 | 2ax7 | HUMAN | N/A |
| P10275 | 2ax8 | HUMAN | Identical to ligand 2ax7 |
| P10275 | 2ax9 | HUMAN | N/A |
| P10275 | 2axa | HUMAN | N/A |
| P10275 | 2hvc | HUMAN | N/A |
| P15207 | 2ihq | RAT | N/A |
| P15207 | 2nw4 | RAT | N/A |
| P10275 | 2oz7 | HUMAN | N/A |
| P10275 | 2pio | HUMAN | Identical ligand to 1i37 |
| P10275 | 2pip | HUMAN | Identical ligand to 1i37 |

| UniProt Code | PDB Code | Organism | Reason for Rejection |
|---|---|---|---|
| P10275 | 2piq | HUMAN | Identical ligand to 1i37 |
| P10275 | 2pir | HUMAN | Identical ligand to 1i37 |
| P10275 | 2pit | HUMAN | Identical ligand to 1i37 |
| P10275 | 2piu | HUMAN | Identical ligand to 1i37 |
| P10275 | 2piv | HUMAN | Identical ligand to 1i37 |
| P10275 | 2piw | HUMAN | Identical ligand to 1i37 |
| P10275 | 2pix | HUMAN | Identical ligand to 1i37 |
| P10275 | 2pkl | HUMAN | Identical ligand to 1i37 |
| P10275 | 2pnu | HUMAN | N/A |
| P10275 | 2q7i | HUMAN | Identical ligand to 1i37 |
| P10275 | 2q7j | HUMAN | Identical ligand to 1i37 |
| P10275 | 2q7k | HUMAN | Identical ligand to 1i37 |
| P10275 | 2q7l | HUMAN | Identical ligand to 1i37 |
| P19091 | 2qpy | MOUSE | Identical ligand to 1i37 |
| P10275 | 2z4j | HUMAN | Identical ligand to 1i37 |
| P10275 | 3b5r | HUMAN | N/A |
| P10275 | 3b65 | HUMAN | N/A |
| P10275 | 3b66 | HUMAN | N/A |
| P10275 | 3b67 | HUMAN | N/A |
| P10275 | 3b68 | HUMAN | N/A |
| P15207 | 3gow | RAT | N/A |
| P10275 | 3l3x | HUMAN | Identical ligand to 1i37 |
| P10275 | 3l3z | HUMAN | Identical ligand to 1i37 |

Table B.3.2 PDB structures for use in the ANDR dataset.

| UniProt Code | PDB Code | Organism | Reason for Rejection |
|---|---|---|---|
| P56817 | 3hob | HUMAN | N/A |
| P56817 | 2ohs | HUMAN | N/A |
| P56817 | 3bug | HUMAN | N/A |
| P56817 | 3l5d | HUMAN | N/A |
| P56817 | 3inh | HUMAN | N/A |
| P56817 | 3kno | HUMAN | N/A |
| P56817 | 3ine | HUMAN | N/A |
| P56817 | 2zjn | HUMAN | N/A |
| P56817 | 3buf | HUMAN | N/A |
| P56817 | 2zjl | HUMAN | N/A |
| P56817 | 3msl | HUMAN | N/A |
| P56817 | 3ind | HUMAN | N/A |
| P56817 | 2zdz | HUMAN | N/A |
| P56817 | 3buh | HUMAN | N/A |
| P56817 | 3in4 | HUMAN | N/A |

| UniProt Code | PDB Code | Organism | Reason for Rejection |
|---|---|---|---|
| P56817 | 2ohr | HUMAN | N/A |
| P56817 | 3lpi | HUMAN | N/A |
| P56817 | 3l5c | HUMAN | N/A |
| P56817 | 3in3 | HUMAN | N/A |
| P56817 | 3igb | HUMAN | N/A |
| P56817 | 3inf | HUMAN | N/A |
| P56817 | 2ohn | HUMAN | N/A |
| P56817 | 3cib | HUMAN | N/A |
| P56817 | 3ckp | HUMAN | N/A |
| P56817 | 2qu3 | HUMAN | N/A |
| P56817 | 2oht | HUMAN | N/A |
| P56817 | 2wjo | HUMAN | N/A |
| P56817 | 3bra | HUMAN | N/A |
| P56817 | 3l38 | HUMAN | N/A |
| P56817 | 3lpk | HUMAN | N/A |
| P56817 | 3l3a | HUMAN | N/A |
| P56817 | 3hw1 | HUMAN | N/A |
| P56817 | 2qu2 | HUMAN | N/A |
| P56817 | 2ze1 | HUMAN | N/A |
| P56817 | 2ohu | HUMAN | N/A |
| P56817 | 2ohp | HUMAN | N/A |
| P56817 | 3kmy | HUMAN | N/A |
| P56817 | 2ohm | HUMAN | N/A |
| P56817 | 2va5 | HUMAN | N/A |
| P56817 | 2ofo | HUMAN | N/A |
| P56817 | 2qmd | HUMAN | N/A |
| P56817 | 3lhg | HUMAN | N/A |
| P56817 | 3l5f | HUMAN | N/A |
| P56817 | 3msk | HUMAN | N/A |
| P56817 | 2va6 | HUMAN | N/A |
| P56817 | 2va7 | HUMAN | N/A |
| P56817 | 2ohl | HUMAN | N/A |
| P56817 | 2q11 | HUMAN | N/A |
| P56817 | 3l5e | HUMAN | N/A |
| P56817 | 3cic | HUMAN | N/A |
| P56817 | 2ohq | HUMAN | N/A |
| P56817 | 2ohk | HUMAN | N/A |
| P56817 | 3lnk | HUMAN | N/A |
| P56817 | 3msj | HUMAN | N/A |
| P56817 | 2q15 | HUMAN | N/A |
| P56817 | 3l59 | HUMAN | N/A |
| P56817 | 3ckr | HUMAN | N/A |
| P56817 | 2qmf | HUMAN | N/A |
| P56817 | 3cid | HUMAN | N/A |
| P56817 | 1xs7 | HUMAN | N/A |

| UniProt Code | PDB Code | Organism | Reason for Rejection |
|---|---|---|---|
| P56817 | 3kmx | HUMAN | N/A |
| P56817 | 3k5c | HUMAN | N/A |
| P56817 | 2zji | HUMAN | N/A |
| P56817 | 2qp8 | HUMAN | N/A |
| P56817 | 3ivh | HUMAN | N/A |
| P56817 | 2iqg | HUMAN | N/A |
| P56817 | 3exo | HUMAN | N/A |
| P56817 | 2hm1 | HUMAN | N/A |
| P56817 | 3lpj | HUMAN | N/A |
| P56817 | 3ivi | HUMAN | N/A |
| P56817 | 2qmg | HUMAN | N/A |
| P56817 | 3fkt | HUMAN | N/A |
| P56817 | 3kyr | HUMAN | N/A |
| P56817 | 2p83 | HUMAN | N/A |
| P56817 | 2b8v | HUMAN | N/A |
| P56817 | 3k5d | HUMAN | N/A |
| P56817 | 2qk5 | HUMAN | N/A |
| P56817 | 3nsh | HUMAN | N/A |
| P56817 | 3l58 | HUMAN | N/A |
| P56817 | 2p8h | HUMAN | N/A |
| P56817 | 2hiz | HUMAN | N/A |
| P56817 | 3l5b | HUMAN | N/A |
| P56817 | 2ntr | HUMAN | N/A |
| P56817 | 2zjk | HUMAN | N/A |
| P56817 | 3duy | HUMAN | N/A |
| P56817 | 2qzk | HUMAN | N/A |
| P56817 | 1tqf | HUMAN | N/A |
| P56817 | 2wf3 | HUMAN | N/A |
| P56817 | 2zjm | HUMAN | N/A |
| P56817 | 2wf2 | HUMAN | N/A |
| P56817 | 2b8l | HUMAN | N/A |
| P56817 | 1ym4 | HUMAN | N/A |
| P56817 | 3hvg | HUMAN | N/A |
| P56817 | 2p4j | HUMAN | N/A |
| P56817 | 3k5f | HUMAN | N/A |
| P56817 | 3k5g | HUMAN | N/A |
| P56817 | 2wf1 | HUMAN | N/A |
| P56817 | 3dv5 | HUMAN | N/A |
| P56817 | 2vnm | HUMAN | N/A |
| P56817 | 2wez | HUMAN | N/A |
| P56817 | 2xfk | HUMAN | N/A |
| P56817 | 2vnn | HUMAN | N/A |
| P56817 | 2wfo | HUMAN | N/A |
| P56817 | 2vkm | HUMAN | N/A |
| P56817 | 2xfi | HUMAN | N/A |

| UniProt Code | PDB Code | Organism | Reason for Rejection |
|---|---|---|---|
| P56817 | 2qzl | HUMAN | N/A |
| P56817 | 2g94 | HUMAN | N/A |
| P56817 | 2vie | HUMAN | N/A |
| P56817 | 2zjh | HUMAN | N/A |
| P56817 | 2f3f | HUMAN | N/A |
| P56817 | 2viz | HUMAN | N/A |
| P56817 | 2vij | HUMAN | N/A |
| P56817 | 2wf4 | HUMAN | N/A |
| P56817 | 2ph6 | HUMAN | N/A |
| P56817 | 2vj7 | HUMAN | N/A |
| P56817 | 1xn3 | HUMAN | No Ligand (Peptide only) |
| P56817 | 2irz | HUMAN | N/A |
| P56817 | 3dv1 | HUMAN | N/A |
| P56817 | 2fdp | HUMAN | N/A |
| P56817 | 2iso | HUMAN | N/A |
| P56817 | 2f3e | HUMAN | N/A |
| P56817 | 2xfj | HUMAN | N/A |
| P56817 | 1w51 | HUMAN | N/A |
| P56817 | 2oah | HUMAN | N/A |
| P56817 | 2vj6 | HUMAN | N/A |
| P56817 | 2zjj | HUMAN | N/A |
| P56817 | 2vj9 | HUMAN | N/A |
| P56817 | 2viy | HUMAN | N/A |
| P56817 | 1xn2 | HUMAN | N/A |

Table B.3.3 PDB structures for use in the BACE1 dataset.

| UniProt Code | PDB Code | Organism | Reason for Rejection |
|---|---|---|---|
| P19491 | 1ftj | RAT | N/A |
| P19491 | 1ftk | RAT | N/A |
| P19491 | 1ftl | RAT | N/A |
| P19491 | 1ftm | RAT | N/A |
| P19491 | 1fwo | RAT | Identical ligand to 1ftk |
| P19491 | 1gr2 | RAT | Identical ligand to 1ftk |
| P19491 | 1lb8 | RAT | Identical ligand to 1ftm |
| P19491 | 1lb9 | RAT | N/A |
| P19491 | 1lbb | RAT | Identical ligand to 1ftk |
| P19491 | 1lbc | RAT | Ligand not in binding site |
| P19491 | 1m5b | RAT | N/A |
| P19491 | 1m5c | RAT | N/A |
| P19491 | 1m5d | RAT | Identical ligand to 1m5c |
| P19491 | 1m5e | RAT | N/A |
| P19491 | 1m5f | RAT | Identical ligand to 1m5e |
| P19491 | 1mm6 | RAT | N/A |

| UniProt Code | PDB Code | Organism | Reason for Rejection |
|---|---|---|---|
| P19491 | 1mm7 | RAT | Identical ligand to 1mm6 |
| P19491 | 1mqd | RAT | N/A |
| P19491 | 1mqg | RAT | N/A |
| P19491 | 1mqh | RAT | N/A |
| P19491 | 1mqi | RAT | N/A |
| P19491 | 1mqj | RAT | N/A |
| P19491 | 1ms7 | RAT | Identical ligand to 1mqd |
| P19491 | 1mxu | RAT | Identical ligand to 1mqh |
| P19491 | 1my2 | RAT | Identical ligand to 1ftm |
| P19491 | 1my3 | RAT | Identical ligand to 1mqh |
| P19491 | 1my4 | RAT | Identical ligand to 1mqg |
| P19491 | 1not | RAT | N/A |
| P19491 | 1nnk | RAT | N/A |
| P19491 | 1nnp | RAT | Identical ligand to 1nnk |
| P19491 | 1p1n | RAT | Identical ligand to 1ftk |
| P19491 | 1p1o | RAT | Identical ligand to 1mm6 |
| P19491 | 1p1q | RAT | Identical ligand to 1ftm |
| P19491 | 1p1u | RAT | Identical ligand to 1ftm |
| P19491 | 1p1w | RAT | Identical ligand to 1ftm |
| P19491 | 1syh | RAT | N/A |
| P19491 | 1syi | RAT | Identical ligand to 1syi |
| P19491 | 1xhy | RAT | Identical ligand to 1ftk |
| P19491 | 2aix | RAT | N/A |
| P19491 | 2al4 | RAT | Identical ligand to 1mm6 |
| P19491 | 2al5 | RAT | Identical ligand to 1mqi |
| P19491 | 2anj | RAT | Identical ligand to 1ftk |
| P19491 | 2cmo | RAT | N/A |
| P19491 | 2gfe | RAT | No Ligand (only crystal cofactor) |
| P19491 | 2i3v | RAT | No Ligand (only crystal cofactor) |
| P19491 | 2i3w | RAT | No Ligand (only crystal cofactor) |
| P19491 | 2p2a | RAT | No Ligand (only crystal cofactor) |
| P19491 | 2uxa | RAT | No Ligand (only crystal cofactor) |
| P19491 | 3b6q | RAT | No Ligand (only crystal cofactor) |
| P19491 | 3b6t | RAT | Identical ligand to 1mm6 |
| P19491 | 3b6w | RAT | No Ligand (only crystal cofactor) |
| P19491 | 3b7d | RAT | N/A |
| P19491 | 3bbr | RAT | No Ligand (only crystal cofactor) |
| P19491 | 3bft | RAT | N/A |
| P19491 | 3bfu | RAT | N/A |
| P19491 | 3bki | RAT | N/A |
| P19491 | 3dp6 | RAT | No Ligand (only crystal cofactor) |
| P19491 | 3h03 | RAT | N/A |
| P19491 | 3h06 | RAT | N/A |
| P19491 | 3h6t | RAT | No Ligand (only crystal cofactor) |

| UniProt Code | PDB Code | Organism | Reason for Rejection |
|---|---|---|---|
| P19491 | 3h6u | RAT | No Ligand (only crystal cofactor) |
| P19491 | 3h6v | RAT | No Ligand (only crystal cofactor) |
| P19491 | 3h6w | RAT | No Ligand (only crystal cofactor) |
| P19491 | 3ijx | RAT | N/A |
| P19491 | 3ik6 | RAT | No Ligand (only crystal cofactor) |
| P19491 | 3il1 | RAT | No Ligand (only crystal cofactor) |
| P19491 | 3ilt | RAT | No Ligand (only crystal cofactor) |
| P19491 | 3ilu | RAT | No Ligand (only crystal cofactor) |
| P19491 | 3kg2 | RAT | N/A |
| P19491 | 3kgc | RAT | N/A |
| P19491 | 3lsl | RAT | Ligand not in binding site |
| P19491 | 3o28 | RAT | Ligand not in binding site |
| P19491 | 3o29 | RAT | Ligand not in binding site |
| P19491 | 3o6g | RAT | Ligand not in binding site |
| P19491 | 3o6h | RAT | Ligand not in binding site |
| P19491 | 3o6i | RAT | Ligand not in binding site |
| P19491 | 3pd8 | RAT | N/A |
| P19491 | 3pd9 | RAT | N/A |
| P19491 | 3pmv | RAT | Ligand not in binding site |
| P19491 | 3pmw | RAT | Ligand not in binding site |

Table B.3.4 PDB structures for use in the GRIA2 dataset.

| UniProt Code | PDB Code | Organism | Reason for Rejection |
|---|---|---|---|
| P07900 | 1osf | HUMAN | N/A |
| P07900 | 1uy6 | HUMAN | N/A |
| P07900 | 1uy7 | HUMAN | N/A |
| P07900 | 1uy8 | HUMAN | N/A |
| P07900 | 1uy9 | HUMAN | N/A |
| P07900 | 1uyc | HUMAN | N/A |
| P07900 | 1uyd | HUMAN | N/A |
| P07900 | 1uye | HUMAN | N/A |
| P07900 | 1uyf | HUMAN | N/A |
| P07900 | 1uyg | HUMAN | N/A |
| P07900 | 1uyh | HUMAN | N/A |
| P07900 | 1uyi | HUMAN | N/A |
| P07900 | 1uyk | HUMAN | N/A |
| P07900 | 1yc1 | HUMAN | N/A |
| P07900 | 1yc3 | HUMAN | N/A |
| P07900 | 1yc4 | HUMAN | N/A |
| P07900 | 1yet | HUMAN | N/A |
| P07900 | 2bsm | HUMAN | N/A |

| UniProt Code | PDB Code | Organism | Reason for Rejection |
|---|---|---|---|
| P07900 | 2bto | HUMAN | N/A |
| P07900 | 2byh | HUMAN | N/A |
| P07900 | 2byi | HUMAN | N/A |
| P07900 | 2bz5 | HUMAN | N/A |
| P07900 | 2ccs | HUMAN | N/A |
| P07900 | 2cct | HUMAN | N/A |
| P07900 | 2ccu | HUMAN | N/A |
| P07900 | 2cdd | HUMAN | N/A |
| P07900 | 2fwy | HUMAN | N/A |
| P07900 | 2fwz | HUMAN | N/A |
| P07900 | 2h55 | HUMAN | N/A |
| P07900 | 2jjc | HUMAN | N/A |
| P07900 | 2qf6 | HUMAN | N/A |
| P07900 | 2qfo | HUMAN | N/A |
| P07900 | 2qg0 | HUMAN | N/A |
| P07900 | 2qg2 | HUMAN | N/A |
| P07900 | 2uwd | HUMAN | N/A |
| P07900 | 2vci | HUMAN | N/A |
| P07900 | 2vcj | HUMAN | N/A |
| P07900 | 2wi1 | HUMAN | N/A |
| P07900 | 2wi2 | HUMAN | N/A |
| P07900 | 2wi3 | HUMAN | N/A |
| P07900 | 2wi4 | HUMAN | N/A |
| P07900 | 2wi5 | HUMAN | N/A |
| P07900 | 2wi6 | HUMAN | N/A |
| P07900 | 2wi7 | HUMAN | N/A |
| P07900 | 2xab | HUMAN | N/A |
| P07900 | 2xdk | HUMAN | N/A |
| P07900 | 2xdl | HUMAN | N/A |
| P07900 | 2xds | HUMAN | N/A |
| P07900 | 2xdu | HUMAN | N/A |
| P07900 | 2xdx | HUMAN | N/A |
| P07900 | 2xhr | HUMAN | N/A |
| P07900 | 2xht | HUMAN | N/A |
| P07900 | 2xhx | HUMAN | N/A |
| P07900 | 2xjg | HUMAN | N/A |
| P07900 | 2xjj | HUMAN | N/A |
| P07900 | 2xjx | HUMAN | N/A |
| P07900 | 2xk2 | HUMAN | N/A |
| P07900 | 3bm9 | HUMAN | N/A |
| P07900 | 3bmy | HUMAN | N/A |
| P07900 | 3dob | HUMAN | N/A |

| UniProt Code | PDB Code | Organism | Reason for Rejection |
|---|---|---|---|
| P07900 | 3eko | HUMAN | N/A |
| P07900 | 3ekr | HUMAN | N/A |
| P07900 | 3ft5 | HUMAN | N/A |
| P07900 | 3ft8 | HUMAN | N/A |
| P07900 | 3hek | HUMAN | N/A |
| P07900 | 3hhu | HUMAN | N/A |
| P07900 | 3hyy | HUMAN | N/A |
| P07900 | 3hyz | HUMAN | N/A |
| P07900 | 3hz1 | HUMAN | N/A |
| P07900 | 3hz5 | HUMAN | N/A |
| P07900 | 3inw | HUMAN | N/A |
| P07900 | 3inx | HUMAN | N/A |
| P07900 | 3k97 | HUMAN | N/A |
| P07900 | 3k98 | HUMAN | N/A |
| P07900 | 3k99 | HUMAN | N/A |
| P07900 | 3mnr | HUMAN | N/A |

Table B.3.5 PDB structures for use in the HSP90A dataset.

| UniProt Code | PDB Code | Organism | Reason for Rejection |
|---|---|---|---|
| P00489 | 1a8i | RABBIT | N/A |
| P00489 | 1abb | RABBIT | No Ligand (only cofactor in binding site) |
| P00489 | 1axr | RABBIT | No Ligand (only cofactor in binding site) |
| P00489 | 1b4d | RABBIT | N/A |
| P00489 | 1c50 | RABBIT | N/A |
| P00489 | 1c8k | RABBIT | Identical ligand to 1c50 |
| P00489 | 1c8l | RABBIT | N/A |
| P00489 | 1e1y | RABBIT | Identical ligand to 1c50 |
| P00489 | 1fs4 | RABBIT | Identical ligand to 1b4d |
| P00489 | 1ftq | RABBIT | N/A |
| P00489 | 1ftw | RABBIT | N/A |
| P00489 | 1fty | RABBIT | N/A |
| P00489 | 1fu4 | RABBIT | N/A |
| P00489 | 1fu7 | RABBIT | N/A |
| P00489 | 1fu8 | RABBIT | N/A |
| P00489 | 1gfz | RABBIT | No Ligand (only cofactor in binding site) |
| P00489 | 1gg8 | RABBIT | N/A |
| P00489 | 1ggn | RABBIT | Identical ligand to 1a8i |
| P00489 | 1gpy | RABBIT | No Ligand (only cofactor in binding site) |
| P00489 | 1h5u | RABBIT | N/A |

| UniProt Code | PDB Code | Organism | Reason for Rejection |
|---|---|---|---|
| P00489 | 1hlf | RABBIT | N/A |
| P00489 | 1k06 | RABBIT | Identical ligand to 2qnb |
| P00489 | 1k08 | RABBIT | Identical ligand to 2qnb |
| P00489 | 1kti | RABBIT | N/A |
| P00489 | 1lwn | RABBIT | Identical ligand to 1h5u |
| P00489 | 1lwo | RABBIT | Identical ligand to 1h5u |
| P00489 | 1noi | RABBIT | N/A |
| P00489 | 1noj | RABBIT | Identical ligand to 1noi |
| P00489 | 1nok | RABBIT | Identical ligand to 1noi |
| P00489 | 1p4g | RABBIT | N/A |
| P00489 | 1p4h | RABBIT | Identical ligand to 1fu8 |
| P00489 | 1p4j | RABBIT | N/A |
| P00489 | 1wut | RABBIT | No Ligand (only cofactor in binding site) |
| P00489 | 1wuy | RABBIT | No Ligand (only cofactor in binding site) |
| P00489 | 1wv0 | RABBIT | No Ligand (only cofactor in binding site) |
| P00489 | 1wv1 | RABBIT | No Ligand (only cofactor in binding site) |
| P00489 | 1ww2 | RABBIT | N/A |
| P00489 | 1ww3 | RABBIT | N/A |
| P00489 | 1xc7 | RABBIT | N/A |
| P00489 | 1xkx | RABBIT | N/A |
| P00489 | 1xl0 | RABBIT | N/A |
| P00489 | 1xl1 | RABBIT | N/A |
| P00489 | 1z62 | RABBIT | N/A |
| P00489 | 1z6p | RABBIT | No Ligand (only cofactor in binding site) |
| P00489 | 1z6q | RABBIT | N/A |
| P11217 | 1z8d | HUMAN | N/A |
| P00489 | 2amv | RABBIT | No Ligand (only cofactor in binding site) |
| P00489 | 2f3p | RABBIT | N/A |
| P00489 | 2f3q | RABBIT | N/A |
| P00489 | 2f3s | RABBIT | N/A |
| P00489 | 2f3u | RABBIT | N/A |
| P00489 | 2fet | RABBIT | N/A |
| P00489 | 2ff5 | RABBIT | Identical ligand to 2fet |
| P00489 | 2ffr | RABBIT | N/A |
| P00489 | 2g9q | RABBIT | N/A |
| P00489 | 2g9r | RABBIT | N/A |
| P00489 | 2g9u | RABBIT | Identical ligand to 2g9r |
| P00489 | 2g9v | RABBIT | N/A |

| UniProt Code | PDB Code | Organism | Reason for Rejection |
|---|---|---|---|
| P00489 | 2gpa | RABBIT | Partial protein structure |
| P00489 | 2gpb | RABBIT | Partial protein structure |
| P00489 | 2off | RABBIT | No Ligand (only cofactor in binding site) |
| P00489 | 2pri | RABBIT | No Ligand (only cofactor in binding site) |
| P00489 | 2prj | RABBIT | N/A |
| P00489 | 2pyd | RABBIT | Identical ligand to 1h5u |
| P00489 | 2pyi | RABBIT | N/A |
| P00489 | 2qlm | RABBIT | Identical ligand to 2qlm |
| P00489 | 2qln | RABBIT | N/A |
| P00489 | 2qn1 | RABBIT | No Ligand (only cofactor in binding site) |
| P00489 | 2qn2 | RABBIT | No Ligand (only cofactor in binding site) |
| P00489 | 2qn3 | RABBIT | N/A |
| P00489 | 2qn7 | RABBIT | No Ligand (only cofactor in binding site) |
| P00489 | 2qn8 | RABBIT | N/A |
| P00489 | 2qn9 | RABBIT | N/A |
| P00489 | 2qnb | RABBIT | N/A |
| P00489 | 2qrg | RABBIT | N/A |
| P00489 | 2qrh | RABBIT | N/A |
| P00489 | 2qrm | RABBIT | N/A |
| P00489 | 2qrp | RABBIT | N/A |
| P00489 | 2qrq | RABBIT | N/A |
| P00489 | 2skd | RABBIT | Identical ligand to 1h5u |
| P00489 | 3amv | RABBIT | Identical ligand to 1h5u |
| P00489 | 3bcr | RABBIT | N/A |
| P00489 | 3bcs | RABBIT | N/A |
| P00489 | 3bcu | RABBIT | N/A |
| P00489 | 3bd6 | RABBIT | No Ligand (only cofactor in binding site) |
| P00489 | 3bd7 | RABBIT | N/A |
| P00489 | 3bd8 | RABBIT | N/A |
| P00489 | 3bda | RABBIT | N/A |
| P00489 | 3cut | RABBIT | N/A |
| P00489 | 3cuu | RABBIT | N/A |
| P00489 | 3cuv | RABBIT | N/A |
| P00489 | 3cuw | RABBIT | N/A |
| P00489 | 3e30 | RABBIT | Completely different protein structure – could not align |
| P00489 | 3ebo | RABBIT | N/A |
| P00489 | 3ebp | RABBIT | N/A |

| UniProt Code | PDB Code | Organism | Reason for Rejection |
|---|---|---|---|
| P00489 | 3g2h | RABBIT | N/A |
| P00489 | 3g2i | RABBIT | N/A |
| P00489 | 3g2j | RABBIT | N/A |
| P00489 | 3g2k | RABBIT | N/A |
| P00489 | 3g2l | RABBIT | N/A |
| P00489 | 3g2n | RABBIT | N/A |
| P00489 | 3gpb | RABBIT | N/A |
| P00489 | 3l79 | RABBIT | N/A |
| P00489 | 3l7a | RABBIT | N/A |
| P00489 | 3l7b | RABBIT | N/A |
| P00489 | 3l7c | RABBIT | N/A |
| P00489 | 3l7d | RABBIT | N/A |
| P00489 | 3mqf | RABBIT | N/A |
| P00489 | 3mrt | RABBIT | N/A |
| P00489 | 3mrv | RABBIT | N/A |
| P00489 | 3mrx | RABBIT | N/A |
| P00489 | 3ms2 | RABBIT | N/A |
| P00489 | 3ms4 | RABBIT | N/A |
| P00489 | 3ms7 | RABBIT | N/A |
| P00489 | 3msc | RABBIT | N/A |
| P00489 | 3mt7 | RABBIT | N/A |
| P00489 | 3mt8 | RABBIT | N/A |
| P00489 | 3mt9 | RABBIT | N/A |
| P00489 | 3mtb | RABBIT | N/A |
| P00489 | 3mtd | RABBIT | N/A |
| P00489 | 3nc4 | RABBIT | N/A |
| P00489 | 4gpb | RABBIT | N/A |
| P00489 | 5gpb | RABBIT | N/A |
| P00489 | 6gpb | RABBIT | N/A |
| P00489 | 7gpb | RABBIT | No Ligand (only cofactor in binding site) |

Table B.3.6 PDB structures for use in the PYGM dataset.

| UniProt Code | PDB Code | Organism | Reason for Rejection |
|---|---|---|---|
| P24941 | 1aq1 | HUMAN | N/A |
| P24941 | 1b38 | HUMAN | N/A |
| P24941 | 1di8 | HUMAN | N/A |
| P24941 | 1dm2 | HUMAN | N/A |
| P24941 | 1e1v | HUMAN | N/A |
| P24941 | 1e1x | HUMAN | N/A |

| UniProt Code | PDB Code | Organism | Reason for Rejection |
|---|---|---|---|
| P24941 | 1e9h | HUMAN | N/A |
| P24941 | 1fvt | HUMAN | N/A |
| P24941 | 1fvv | HUMAN | N/A |
| P24941 | 1g5s | HUMAN | N/A |
| P24941 | 1gih | HUMAN | N/A |
| P24941 | 1gii | HUMAN | Identical ligand to 1gih |
| P24941 | 1gij | HUMAN | N/A |
| P24941 | 1gz8 | HUMAN | N/A |
| P24941 | 1h00 | HUMAN | N/A |
| P24941 | 1h01 | HUMAN | N/A |
| P24941 | 1h07 | HUMAN | N/A |
| P24941 | 1h08 | HUMAN | N/A |
| P24941 | 1h0v | HUMAN | N/A |
| P24941 | 1h0w | HUMAN | N/A |
| P24941 | 1h1p | HUMAN | N/A |
| P24941 | 1h1q | HUMAN | N/A |
| P24941 | 1h1s | HUMAN | N/A |
| P24941 | 1hck | HUMAN | Identical ligand to 1b38 |
| P24941 | 1jsv | HUMAN | N/A |
| P24941 | 1ke5 | HUMAN | N/A |
| P24941 | 1ke6 | HUMAN | N/A |
| P24941 | 1ke7 | HUMAN | N/A |
| P24941 | 1ke8 | HUMAN | N/A |
| P24941 | 1ke9 | HUMAN | N/A |
| P24941 | 1ogu | HUMAN | N/A |
| P24941 | 1oi9 | HUMAN | N/A |
| P24941 | 1oiq | HUMAN | N/A |
| P24941 | 1oir | HUMAN | N/A |
| P24941 | 1oit | HUMAN | N/A |
| P24941 | 1oiu | HUMAN | N/A |
| P24941 | 1oiy | HUMAN | N/A |
| P24941 | 1ol2 | HUMAN | No Ligand |
| P24941 | 1p2a | HUMAN | N/A |
| P24941 | 1p5e | HUMAN | N/A |
| P24941 | 1pf8 | HUMAN | N/A |
| P24941 | 1pkd | HUMAN | N/A |
| P24941 | 1pxi | HUMAN | N/A |
| P24941 | 1pxj | HUMAN | N/A |
| P24941 | 1pxk | HUMAN | N/A |
| P24941 | 1pxl | HUMAN | N/A |
| P24941 | 1pxm | HUMAN | N/A |
| P24941 | 1pxn | HUMAN | N/A |

| UniProt Code | PDB Code | Organism | Reason for Rejection |
|---|---|---|---|
| P24941 | 1pxp | HUMAN | N/A |
| P24941 | 1pye | HUMAN | N/A |
| P24941 | 1r78 | HUMAN | N/A |
| P24941 | 1urw | HUMAN | N/A |
| P24941 | 1v1k | HUMAN | N/A |
| P24941 | 1vyw | HUMAN | N/A |
| P24941 | 1vyz | HUMAN | N/A |
| P24941 | 1wox | HUMAN | N/A |
| P24941 | 1w8c | HUMAN | N/A |
| P24941 | 1wcc | HUMAN | N/A |
| P24941 | 1y8y | HUMAN | N/A |
| P24941 | 1y91 | HUMAN | N/A |
| P24941 | 1ykr | HUMAN | N/A |
| P24941 | 2a0c | HUMAN | N/A |
| P24941 | 2a4l | HUMAN | N/A |
| P24941 | 2b52 | HUMAN | N/A |
| P24941 | 2b53 | HUMAN | N/A |
| P24941 | 2b54 | HUMAN | N/A |
| P24941 | 2b55 | HUMAN | N/A |
| P24941 | 2bhe | HUMAN | N/A |
| P24941 | 2bhh | HUMAN | N/A |
| P24941 | 2bkz | HUMAN | N/A |
| P24941 | 2bpm | HUMAN | N/A |
| P24941 | 2btr | HUMAN | N/A |
| P24941 | 2bts | HUMAN | N/A |
| P24941 | 2c4g | HUMAN | N/A |
| P24941 | 2c5v | HUMAN | N/A |
| P24941 | 2c5x | HUMAN | N/A |
| P24941 | 2c5y | HUMAN | N/A |
| P24941 | 2c68 | HUMAN | N/A |
| P24941 | 2c69 | HUMAN | N/A |
| P24941 | 2c6i | HUMAN | N/A |
| P24941 | 2c6k | HUMAN | N/A |
| P24941 | 2c6l | HUMAN | N/A |
| P24941 | 2c6m | HUMAN | N/A |
| P24941 | 2c6o | HUMAN | N/A |
| P24941 | 2c6t | HUMAN | N/A |
| P24941 | 2cjm | HUMAN | N/A |
| P24941 | 2clx | HUMAN | N/A |
| P24941 | 2duv | HUMAN | N/A |
| P24941 | 2exm | HUMAN | N/A |
| P24941 | 2fvd | HUMAN | N/A |

| UniProt Code | PDB Code | Organism | Reason for Rejection |
|---|---|---|---|
| P24941 | 2g9x | HUMAN | N/A |
| P24941 | 2i40 | HUMAN | N/A |
| P24941 | 2iw6 | HUMAN | N/A |
| P24941 | 2iw8 | HUMAN | Identical ligand to 1h1q |
| P24941 | 2iw9 | HUMAN | Identical ligand to 1h1q |
| P24941 | 2r3h | HUMAN | N/A |
| P24941 | 2r64 | HUMAN | N/A |
| P24941 | 2uue | HUMAN | N/A |
| P24941 | 2uzb | HUMAN | N/A |
| P24941 | 2uzd | HUMAN | N/A |
| P24941 | 2uze | HUMAN | N/A |
| P24941 | 2uzl | HUMAN | N/A |
| P24941 | 2uzn | HUMAN | N/A |
| P24941 | 2uzo | HUMAN | N/A |
| P24941 | 2vod | HUMAN | N/A |
| P24941 | 2vta | HUMAN | N/A |
| P24941 | 2vth | HUMAN | N/A |
| P24941 | 2vti | HUMAN | N/A |
| P24941 | 2vtj | HUMAN | N/A |
| P24941 | 2vtl | HUMAN | N/A |
| P24941 | 2vtm | HUMAN | N/A |
| P24941 | 2vtn | HUMAN | N/A |
| P24941 | 2vto | HUMAN | N/A |
| P24941 | 2vtp | HUMAN | N/A |
| P24941 | 2vtq | HUMAN | N/A |
| P24941 | 2vtr | HUMAN | N/A |
| P24941 | 2vts | HUMAN | N/A |
| P24941 | 2vtt | HUMAN | N/A |
| P24941 | 2vu3 | HUMAN | N/A |
| P24941 | 2vv9 | HUMAN | N/A |
| P24941 | 2w05 | HUMAN | N/A |
| P24941 | 2w06 | HUMAN | N/A |
| P24941 | 2w17 | HUMAN | N/A |
| P24941 | 2w1h | HUMAN | N/A |
| P24941 | 2wev | HUMAN | N/A |
| P24941 | 2wih | HUMAN | N/A |
| P24941 | 2wip | HUMAN | N/A |
| P24941 | 2wpa | HUMAN | N/A |
| P24941 | 2wxv | HUMAN | N/A |
| P24941 | 2x1n | HUMAN | N/A |
| P24941 | 3bht | HUMAN | N/A |
| P24941 | 3bhu | HUMAN | N/A |

| UniProt Code | PDB Code | Organism | Reason for Rejection |
|---|---|---|---|
| P24941 | 3bhv | HUMAN | N/A |
| P24941 | 3ddp | HUMAN | N/A |
| P24941 | 3ddq | HUMAN | N/A |
| P24941 | 3dog | HUMAN | N/A |
| P24941 | 3eid | HUMAN | N/A |
| P24941 | 3ej1 | HUMAN | N/A |
| P24941 | 3eoc | HUMAN | N/A |
| P24941 | 3ezr | HUMAN | N/A |
| P24941 | 3ezv | HUMAN | N/A |
| P24941 | 3f5x | HUMAN | N/A |
| P24941 | 3fz1 | HUMAN | N/A |
| P24941 | 3ig7 | HUMAN | N/A |
| P24941 | 3igg | HUMAN | N/A |
| P24941 | 3le6 | HUMAN | N/A |
| P24941 | 3lfn | HUMAN | N/A |
| P24941 | 3lfq | HUMAN | N/A |
| P24941 | 3lfs | HUMAN | N/A |
| P24941 | 3my5 | HUMAN | N/A |
| P24941 | 3ns9 | HUMAN | N/A |
| P24941 | 3pxf | HUMAN | N/A |
| P24941 | 3pxq | HUMAN | N/A |
| P24941 | 3pxy | HUMAN | N/A |
| P24941 | 3pxz | HUMAN | N/A |
| P24941 | 3py0 | HUMAN | Identical ligand to 1pf8 |
| P24941 | 3py1 | HUMAN | Identical ligand to 1pf8 |

Table B.3.7 PDB structures for use in the CDK2 dataset.

| UniProt Code | PDB Code | Organism | Reason for Rejection |
|---|---|---|---|
| P00374 | 1boz | HUMAN | N/A |
| P00374 | 1dfr | HUMAN | N/A |
| P00374 | 1dhf | HUMAN | N/A |
| P00374 | 1dlr | HUMAN | N/A |
| P00374 | 1dls | HUMAN | N/A |
| P00374 | 1hfp | HUMAN | N/A |
| P00374 | 1hfq | HUMAN | Identical ligand to 1dlr |
| P00374 | 1hfr | HUMAN | Identical ligand to 1dlr |
| P00374 | 1kms | HUMAN | N/A |
| P00374 | 1kmv | HUMAN | N/A |

| UniProt Code | PDB Code | Organism | Reason for Rejection |
|---|---|---|---|
| P00374 | 1mvs | HUMAN | N/A |
| P00374 | 1mvt | HUMAN | N/A |
| P00374 | 1ohj | HUMAN | N/A |
| P00374 | 1ohk | HUMAN | N/A |
| P00374 | 1pd8 | HUMAN | N/A |
| P00374 | 1pd9 | HUMAN | Identical ligand to 1pd8 |
| P00374 | 1pdb | HUMAN | No Ligand |
| P00374 | 1s3u | HUMAN | N/A |
| P00374 | 1s3v | HUMAN | Identical ligand to 1s3u |
| P00374 | 1s3w | HUMAN | N/A |
| P00374 | 1u71 | HUMAN | N/A |
| P00374 | 1u72 | HUMAN | N/A |
| P00374 | 1yho | HUMAN | N/A |
| P00374 | 2c2s | HUMAN | N/A |
| P00374 | 2c2t | HUMAN | N/A |
| P00374 | 2dhf | HUMAN | N/A |
| P00374 | 2w3a | HUMAN | N/A |
| P00374 | 2w3b | HUMAN | N/A |
| P00374 | 2w3m | HUMAN | N/A |
| P00374 | 3eig | HUMAN | N/A |
| P00374 | 3f8y | HUMAN | N/A |
| P00374 | 3f8z | HUMAN | Identical ligand to 3f8z |
| P00374 | 3f9i | HUMAN | N/A |
| P00374 | 3fs6 | HUMAN | Identical ligand to 3f8z |
| P00374 | 3ghc | HUMAN | N/A |
| P00374 | 3ghv | HUMAN | Identical ligand to 3ghc |
| P00374 | 3ghw | HUMAN | Identical ligand to 3ghc |
| P00374 | 3gi2 | HUMAN | N/A |
| P00374 | 3gyf | HUMAN | N/A |
| P00374 | 3l3r | HUMAN | N/A |
| P00374 | 3noh | HUMAN | N/A |
| P00374 | 3ntz | HUMAN | N/A |
| P00374 | 3nuo | HUMAN | N/A |
| P00374 | 3nxo | HUMAN | N/A |
| P00374 | 3nxr | HUMAN | N/A |
| P00374 | 3nxt | HUMAN | N/A |
| P00374 | 3nxv | HUMAN | N/A |
| P00374 | 3nxx | HUMAN | N/A |
| P00374 | 3nxy | HUMAN | N/A |
| P00374 | 3nzd | HUMAN | N/A |
| P00374 | 3oaf | HUMAN | N/A |
| P00374 | 3s3v | HUMAN | Identical ligand to 3noH |

| UniProt Code | PDB Code | Organism | Reason for Rejection |
|---|---|---|---|
| P00374 | 3s7a | HUMAN | N/A |
| P00374 | 4ddr | HUMAN | N/A |
| P00374 | 4g95 | HUMAN | N/A |
| P00374 | 4kak | HUMAN | N/A |
| P00374 | 4kbn | HUMAN | N/A |
| P00374 | 4kd7 | HUMAN | N/A |
| P00374 | 4keb | HUMAN | N/A |
| P00374 | 4kfj | HUMAN | N/A |
| P00374 | 4m6j | HUMAN | No Ligand |
| P00374 | 4m6k | HUMAN | N/A |
| P00374 | 4m6l | HUMAN | N/A |
| P00374 | 4qhv | HUMAN | N/A |
| P00374 | 4qjc | HUMAN | N/A |
| P00374 | 5hpb | HUMAN | N/A |
| P00374 | 5hqy | HUMAN | N/A |
| P00374 | 5hqz | HUMAN | N/A |
| P00374 | 5hsu | HUMAN | N/A |
| P00374 | 5hsr | HUMAN | Identical ligand to 5hsu |
| P00374 | 5ht4 | HUMAN | N/A |
| P00374 | 5ht5 | HUMAN | N/A |
| P00374 | 5hui | HUMAN | N/A |
| P00374 | 5hvb | HUMAN | N/A |
| P00374 | 5hve | HUMAN | N/A |

Table B.3.8 PDB structures for use in the DHFR dataset.

ADRB2, 5 structures, Run 1

| PDB Code | Ligand Residue Code | Resolution / Å |
|---|---|---|
| 3d4s | TIM | 2.8 |
| 3ny8 | JRZ | 2.84 |
| 3ny9 | JSZ | 2.84 |
| 3nya | JTZ | 3.16 |
| 3pog | PoG | 3.5 |

Table B.3.9 PDB structures used to construct the CRANkS grids for run 1 of 5 structures for target ADRB2

ADRB2, 5 structures, Run 2

| PDB Code | Ligand Residue Code | Resolution / Å |
|---|---|---|
| 2rh1 | CAU | 2.4 |
| 3d4s | TIM | 2.8 |
| 3ny9 | JSZ | 2.84 |
| 3nya | JTZ | 3.16 |
| 3pog | PoG | 3.5 |

Table B.3.10 PDB structures used to construct the CRANkS grids for run 2 of 5 structures for target ADRB2

ADRB2, 5 structures, Run 3

| PDB Code | Ligand Residue Code | Resolution / Å |
|---|---|---|
| 2rh1 | CAU | 2.4 |
| 3ny8 | JRZ | 2.84 |
| 3nya | JTZ | 3.16 |
| 3pog | PoG | 3.5 |
| 3pds | ERC | 3.5 |

Table B.3.11 PDB structures used to construct the CRANkS grids for run 3 of 5 structures for target ADRB2

ADRB2, 5 structures, Run 4

| PDB Code | Ligand Residue Code | Resolution / Å |
|---|---|---|
| 2rh1 | CAU | 2.4 |
| 3d4s | TIM | 2.8 |
| 3ny8 | JRZ | 2.84 |
| 3ny9 | JSZ | 2.84 |
| 3nya | JTZ | 3.16 |

Table B.3.12 PDB structures used to construct the CRANkS grids for run 4 of 5 structures for target ADRB2

ADRB2, 5 structures, Run 5

| PDB Code | Ligand Residue Code | Resolution / Å |
|---|---|---|
| 2rh1 | CAU | 2.4 |
| 3d4s | TIM | 2.8 |
| 3ny8 | JRZ | 2.84 |
| 3pog | PoG | 3.5 |

| | | |
|---|---|---|
| 3pds | ERC | 3.5 |

Table B.3.13 PDB structures used to construct the CRANkS grids for run 5 of 5 structures for target ADRB2

ANDR, 5 structures, Run 1

| PDB Code | Ligand Residue Code | Resolution / Å |
|---|---|---|
| 2amb | 17H | 1.75 |
| 2ax9 | BHM | 1.65 |
| 2hvc | LGD | 2.10 |
| 3b5r | B5R | 1.80 |
| 3b66 | B66 | 1.65 |

Table B.3.14 PDB structures used to construct the CRANkS grids for run 1 of 5 structures for target ANDR

ANDR, 5 structures, Run 2

| PDB Code | Ligand Residue Code | Resolution / Å |
|---|---|---|
| 1i37 | DHT | 2.00 |
| 2amb | 17H | 1.75 |
| 2axa | FHM | 1.80 |
| 3b5r | B5R | 1.80 |
| 3b66 | B66 | 1.65 |

Table B.3.15 PDB structures used to construct the CRANkS grids for run 2 of 5 structures for target ANDR

ANDR, 5 structures, Run 3

| PDB Code | Ligand Residue Code | Resolution / Å |
|---|---|---|
| 1gs4 | ZK5 | 1.95 |
| 3b5r | B5R | 1.80 |
| 3b66 | B66 | 1.65 |
| 3b68 | B68 | 1.90 |
| 3g0w | LGB | 1.95 |

Table B.3.16 PDB structures used to construct the CRANkS grids for run 3 of 5 structures for target ANDR

ANDR, 5 structures, Run 4

| PDB Code | Ligand Residue Code | Resolution / Å |
|---|---|---|
| 1gs4 | ZK5 | 1.95 |
| 2amb | 17H | 1.75 |
| 2ax7 | FHM | 1.90 |

| | | |
|---|---|---|
| 2hvc | LGD | 2.10 |
| 2nw4 | 8NH | 3.00 |

Table B.3.17 PDB structures used to construct the CRANkS grids for run 4 of 5 structures for target ANDR

ANDR, 5 structures, Run 5

| PDB Code | Ligand Residue Code | Resolution / Å |
|---|---|---|
| 1gs4 | ZK5 | 1.95 |
| 1i37 | DHT | 2.00 |
| 1xnn | HYQ | 2.20 |
| 2amb | 17H | 1.75 |
| 2ax7 | FHM | 1.90 |

Table B.3.18 PDB structures used to construct the CRANkS grids for run 5 of 5 structures for target ANDR

ANDR, 10 structures, Run 1

| PDB Code | Ligand Residue Code | Resolution / Å |
|---|---|---|
| 2amb | 17H | 1.75 |
| 2ao6 | R18 | 1.89 |
| 2ax7 | FHM | 1.90 |
| 2ax9 | BHM | 1.65 |
| 2axa | FHM | 1.80 |
| 2hvc | LGD | 2.10 |
| 2ihq | LG7 | 2.00 |
| 3b5r | B5R | 1.80 |
| 3b65 | 3B6 | 1.80 |
| 3b66 | B66 | 1.65 |

Table B.3.19 PDB structures used to construct the CRANkS grids for run 1 of 10 structures for target ANDR

ANDR, 10 structures, Run 2

| PDB Code | Ligand Residue Code | Resolution / Å |
|---|---|---|
| 1e3g | R18 | 2.40 |
| 1i37 | DHT | 2.00 |
| 2amb | 17H | 1.75 |
| 2ax9 | BHM | 1.65 |
| 2axa | FHM | 1.80 |
| 2oz7 | CA4 | 1.80 |
| 2pnu | ENM | 1.65 |
| 3b5r | B5R | 1.80 |
| 3b65 | 3B6 | 1.80 |

| | | |
|---|---|---|
| 3b66 | B66 | 1.65 |

Table B.3.20 PDB structures used to construct the CRANkS grids for run 2 of 10 structures for target ANDR

ANDR, 10 structures, Run 3

| PDB Code | Ligand Residue Code | Resolution / Å |
|---|---|---|
| 1gs4 | ZK5 | 1.95 |
| 1xnn | HYQ | 2.20 |
| 2a06 | R18 | 1.89 |
| 2ihq | LG7 | 2.00 |
| 2nw4 | 8NH | 3.00 |
| 2oz7 | CA4 | 1.80 |
| 3b5r | B5R | 1.80 |
| 3b66 | B66 | 1.65 |
| 3b68 | B68 | 1.90 |
| 3g0w | LGB | 1.95 |

Table B.3.21 PDB structures used to construct the CRANkS grids for run 3 of 10 structures for target ANDR

ANDR, 10 structures, Run 4

| PDB Code | Ligand Residue Code | Resolution / Å |
|---|---|---|
| 1e3g | R18 | 2.40 |
| 1gs4 | ZK5 | 1.95 |
| 2amb | 17H | 1.75 |
| 2ax7 | FHM | 1.90 |
| 2axa | FHM | 1.80 |
| 2hvc | LGD | 2.10 |
| 2nw4 | 8NH | 3.00 |
| 2oz7 | CA4 | 1.80 |
| 3b66 | B66 | 1.65 |
| 3g0w | LGB | 1.95 |

Table B.3.22 PDB structures used to construct the CRANkS grids for run 4 of 10 structures for target ANDR

ANDR, 10 structures, Run 5

| PDB Code | Ligand Residue Code | Resolution / Å |
|---|---|---|
| 1gs4 | ZK5 | 1.95 |
| 1i37 | DHT | 2.00 |
| 1xnn | HYQ | 2.20 |
| 1z95 | 198.00 | 1.80 |

| | | |
|---|---|---|
| 2amb | 17H | 1.75 |
| 2ax7 | FHM | 1.90 |
| 2hvc | LGD | 2.10 |
| 3b5r | B5R | 1.80 |
| 3b65 | 3B6 | 1.80 |
| 3b68 | B68 | 1.90 |

Table B.3.23 PDB structures used to construct the CRANkS grids for run 5 of 10 structures for target ANDR

BACE1, 5 structures, Run 1

| PDB Code | Ligand Residue Code | Resolution / Å |
|---|---|---|
| 2ohq | 7IP | 2.1 |
| 3l5c | BDQ | 1.8 |
| 2ofo | CMZ | 2.25 |
| 1xs7 | MMI | 2.8 |
| 2vj6 | VG5 | 1.8 |

Table B.3.24 PDB structures used to construct the CRANkS grids for run 1 of 5 structures for target BACE1

BACE1, 5 structures, Run 2

| PDB Code | Ligand Residue Code | Resolution / Å |
|---|---|---|
| 3lnk | 74A | 1.8 |
| 2vnn | CM7 | 1.87 |
| 1xs7 | MMI | 2.8 |
| 2vj6 | VG5 | 1.8 |

Table B.3.25 PDB structures used to construct the CRANkS grids for run 2 of 5 structures for target BACE1

BACE1, 5 structures, Run 3

| PDB Code | Ligand Residue Code | Resolution / Å |
|---|---|---|
| 2p4j | 23I | 2.5 |
| 2ohl | 2AQ | 2.65 |
| 2viy | VG3 | 1.82 |
| 2g94 | ZPQ | 1.86 |
| 2wfo | ZYo | 1.6 |

Table B.3.26 PDB structures used to construct the CRANkS grids for run 3 of 5 structures for target BACE1

BACE1, 5 structures, Run 4

| PDB Code | Ligand Residue Code | Resolution / Å |
|---|---|---|
| 2ohp | 6IP | 2.25 |
| 2f3e | AXQ | 2.11 |
| 2qmf | CS9 | 1.75 |
| 2zji | F1I | 2.3 |
| 2zjl | F1L | 2.1 |

Table B.3.27 PDB structures used to construct the CRANkS grids for run 4 of 5 structures for target BACE1

BACE1, 5 structures, Run 5

| PDB Code | Ligand Residue Code | Resolution / Å |
|---|---|---|
| 1ym4 | 24M | 2.25 |
| 3cib | 314 | 1.72 |
| 3kno | 3TO | 1.9 |
| 2ohp | 6IP | 2.25 |
| 3dv5 | BAV | 2.1 |

Table B.3.28 PDB structures used to construct the CRANkS grids for run 5 of 5 structures for target BACE1

BACE1, 10 structures, Run 1

| PDB Code | Ligand Residue Code | Resolution / Å |
|---|---|---|
| 2ohq | 7IP | 2.1 |
| 2ohs | 9IP | 2.45 |
| 3l59 | BDJ | 2 |
| 3l5c | BDQ | 1.8 |
| 2vij | C44 | 1.6 |
| 2ofo | CMZ | 2.25 |
| 3msk | EV4 | 2 |
| 1xs7 | MMI | 2.8 |
| 2oah | QIN | 1.8 |
| 2vj6 | VG5 | 1.8 |

Table B.3.29 PDB structures used to construct the CRANkS grids for run 1 of 10 structures for target BACE1

BACE1, 10 structures, Run 2

| PDB Code | Ligand Residue Code | Resolution / Å |
|---|---|---|
| 3k5c | 0BI | 2.12 |
| 2zdz | 310 | 2 |
| 2b8v | 3BN | 1.8 |
| 3lnk | 74A | 1.8 |
| 3k5g | BJC | 2 |
| 2vnn | CM7 | 1.87 |
| 2iqg | F2I | 1.7 |
| 1xs7 | MMI | 2.8 |
| 2oah | QIN | 1.8 |
| 2vj6 | VG5 | 1.8 |

Table B.3.30 PDB structures used to construct the CRANkS grids for run 2 of 10 structures for target BACE1

BACE1, 10 structures, Run 3

| PDB Code | Ligand Residue Code | Resolution / Å |
|---|---|---|
| 2p4j | 23I | 2.5 |
| 2ohl | 2AQ | 2.65 |

| | | |
|---|---|---|
| 2xfk | AA9 | 1.8 |
| 3msk | EV4 | 2 |
| 2zji | F1I | 2.3 |
| 3kmx | G00 | 1.7 |
| 1w51 | L01 | 2.55 |
| 2viy | VG3 | 1.82 |
| 2g94 | ZPQ | 1.86 |
| 2wfo | ZY0 | 1.6 |

Table B.3.31 PDB structures used to construct the CRANkS grids for run 3 of 10 structures for target BACE1

BACE1, 10 structures, Run 4

| PDB Code | Ligand Residue Code | Resolution / Å |
|---|---|---|
| 3ckp | 12 | 2.3 |
| 1ym4 | 24M | 2.25 |
| 3l3a | 625 | 2.362 |
| 2ohp | 6IP | 2.25 |
| 2ohq | 7IP | 2.1 |
| 2f3e | AXQ | 2.11 |
| 2qmf | CS9 | 1.75 |
| 2zji | F1I | 2.3 |
| 2zjl | F1L | 2.1 |
| 2viz | VG4 | 1.6 |

Table B.3.32 PDB structures used to construct the CRANkS grids for run 4 of 10 structures for target BACE1

BACE1, 10 structures, Run 5

| PDB Code | Ligand Residue Code | Resolution / Å |
|---|---|---|
| 1ym4 | 24M | 2.25 |
| 3cib | 314 | 1.72 |
| 3kno | 3TO | 1.9 |
| 3exo | 5MS | 2.1 |
| 2ohp | 6IP | 2.25 |
| 3dv5 | BAV | 2.1 |
| 3l59 | BDJ | 2 |
| 1xs7 | MMI | 2.8 |
| 2qmg | SC6 | 1.89 |
| 3lpi | Z74 | 2.05 |

Table B.3.33 PDB structures used to construct the CRANkS grids for run 5 of 10 structures for target BACE1

BACE1, 25 structures, Run 1

| PDB Code | Ligand Residue Code | Resolution / Å |
|---|---|---|
| 2zdz | 310 | 2 |
| 3lnk | 74A | 1.8 |
| 2ohq | 7IP | 2.1 |

| | | |
|---|---|---|
| 2ohr | 8IP | 2.25 |
| 2ohs | 9IP | 2.45 |
| 2xfk | AA9 | 1.8 |
| 3l59 | BDJ | 2 |
| 3l5c | BDQ | 1.8 |
| 3l5e | BDW | 1.53 |
| 2va7 | C27 | 2.2 |
| 2vij | C44 | 1.6 |
| 2vnm | CM8 | 1.79 |
| 2ofo | CMZ | 2.25 |
| 3msk | EV4 | 2 |
| 2zjk | F1K | 3 |
| 2irz | I02 | 1.8 |
| 2ntr | L00 | 1.8 |
| 2hm1 | LIQ | 2.2 |
| 1xs7 | MMI | 2.8 |
| 2oah | QIN | 1.8 |
| 2qp8 | SC7 | 1.5 |
| 2vj6 | VG5 | 1.8 |
| 3k5d | XLI | 2.9 |
| 2q11 | XX4 | 2.4 |
| 2wf3 | ZY3 | 2.08 |

Table B.3.34 PDB structures used to construct the CRANkS grids for run 1 of 25 structures for target BACE1

BACE1, 25 structures, Run 2

| PDB Code | Ligand Residue Code | Resolution / Å |
|---|---|---|
| 3ckr | 9 | 2.7 |
| 3k5c | 0BI | 2.12 |
| 2zdz | 310 | 2 |
| 2b8v | 3BN | 1.8 |
| 2b8l | 5HA | 1.7 |
| 3exo | 5MS | 2.1 |
| 3l3a | 625 | 2.362 |
| 3lnk | 74A | 1.8 |
| 3nsh | 957 | 2.2 |
| 3k5f | AYH | 2.25 |
| 3l5d | BDV | 1.75 |
| 3l5e | BDW | 1.53 |
| 3l5f | BDX | 1.7 |
| 3k5g | BJC | 2 |
| 2vkm | BSD | 2.05 |
| 2vnn | CM7 | 1.87 |
| 2qmf | CS9 | 1.75 |
| 2iqg | F2I | 1.7 |
| 2qzl | IXS | 1.8 |
| 1xs7 | MMI | 2.8 |
| 2oah | QIN | 1.8 |

| | | |
|---|---|---|
| 2viy | VG3 | 1.82 |
| 2vj6 | VG5 | 1.8 |
| 3k5d | XLI | 2.9 |
| 3lhg | Z81 | 2.1 |

Table B.3.35 PDB structures used to construct the CRANkS grids for run 2 of 25 structures for target BACE1

BACE1, 25 structures, Run 3

| PDB Code | Ligand Residue Code | Resolution / Å |
|---|---|---|
| 3k5c | 0BI | 2.12 |
| 3ivh | 1LI | 1.8 |
| 2p4j | 23I | 2.5 |
| 2ohl | 2AQ | 2.65 |
| 2ze1 | 4I1 | 2.2 |
| 3l38 | 879 | 2.1 |
| 2xfk | AA9 | 1.8 |
| 3buh | AED | 2.3 |
| 3k5f | AYH | 2.25 |
| 3dv5 | BAV | 2.1 |
| 3l5e | BDW | 1.53 |
| 3l5f | BDX | 1.7 |
| 3in4 | BX2 | 2.3 |
| 3l58 | CS5 | 1.8 |
| 2qmf | CS9 | 1.75 |
| 3msk | EV4 | 2 |
| 2zji | F1I | 2.3 |
| 3kmx | G00 | 1.7 |
| 2qzl | IXS | 1.8 |
| 1w51 | L01 | 2.55 |
| 2viy | VG3 | 1.82 |
| 2xfi | XFI | 1.73 |
| 2g94 | ZPQ | 1.86 |
| 2wfo | ZY0 | 1.6 |
| 2wez | ZYE | 1.7 |

Table B.3.36 PDB structures used to construct the CRANkS grids for run 3 of 25 structures for target BACE1

BACE1, 25 structures, Run 4

| PDB Code | Ligand Residue Code | Resolution / Å |
|---|---|---|
| 3ckp | 12 | 2.3 |
| 3k5c | 0BI | 2.12 |
| 1ym4 | 24M | 2.25 |

| | | |
|---|---|---|
| 3kno | 3TO | 1.9 |
| 3l3a | 625 | 2.362 |
| 2ohp | 6IP | 2.25 |
| 2ohq | 7IP | 2.1 |
| 2ohs | 9IP | 2.45 |
| 2f3e | AXQ | 2.11 |
| 2va7 | C27 | 2.2 |
| 2qk5 | CS5 | 2.2 |
| 2qmf | CS9 | 1.75 |
| 3msl | EV5 | 2.4 |
| 2zji | F1I | 2.3 |
| 2zjl | F1L | 2.1 |
| 2zjm | F1M | 1.9 |
| 3kmx | G00 | 1.7 |
| 2qzl | IXS | 1.8 |
| 1w51 | L01 | 2.55 |
| 1xs7 | MMI | 2.8 |
| 2viz | VG4 | 1.6 |
| 2vj7 | VG6 | 1.6 |
| 2vj9 | VG7 | 1.6 |
| 3ine | X17 | 1.996 |
| 2wez | ZYE | 1.7 |

Table B.3.37 PDB structures used to construct the CRANkS grids for run 4 of 25 structures for target BACE1

BACE1, 25 structures, Run 5

| PDB Code | Ligand Residue Code | Resolution / Å |
|---|---|---|
| 1ym4 | 24M | 2.25 |
| 3ivi | 2LI | 2.2 |
| 3cib | 314 | 1.72 |
| 3kno | 3TO | 1.9 |
| 2qu3 | 462 | 2 |
| 2ohn | 4FP | 2.15 |
| 2b8l | 5HA | 1.7 |
| 3exo | 5MS | 2.1 |
| 2ohp | 6IP | 2.25 |
| 2ohm | 8AP | 2.7 |
| 2xfk | AA9 | 1.8 |
| 3dv5 | BAV | 2.1 |
| 3l59 | BDJ | 2 |
| 2qmd | CS7 | 1.65 |
| 2zji | F1I | 2.3 |
| 2zjl | F1L | 2.1 |
| 2ntr | L00 | 1.8 |
| 1xs7 | MMI | 2.8 |
| 2qmg | SC6 | 1.89 |
| 2qp8 | SC7 | 1.5 |
| 2viy | VG3 | 1.82 |
| 2xfj | VG5 | 1.8 |
| 3ine | X17 | 1.996 |

| | | |
|---|---|---|
| 3lpi | Z74 | 2.05 |
| 3lpj | Z75 | 1.79 |

Table B.3.38 PDB structures used to construct the CRANkS grids for run 5 of 25 structures for target BACE1

BACE1, 50 structures, Run 1

| PDB Code | Ligand Residue Code | Resolution / Å |
|---|---|---|
| 3ckr | 9 | 2.7 |
| 3ckp | 12 | 2.3 |
| 2ohl | 2AQ | 2.65 |
| 2zdz | 310 | 2 |
| 3cib | 314 | 1.72 |
| 2ohn | 4FP | 2.15 |
| 2ohp | 6IP | 2.25 |
| 2ph6 | 712 | 2 |
| 3lnk | 74A | 1.8 |
| 2ohq | 7IP | 2.1 |
| 2ohr | 8IP | 2.25 |
| 2ohs | 9IP | 2.45 |
| 2xfk | AA9 | 1.8 |
| 3buh | AED | 2.3 |
| 3bra | AEF | 2.3 |
| 3l59 | BDJ | 2 |
| 3l5c | BDQ | 1.8 |
| 3l5e | BDW | 1.53 |
| 3k5g | BJC | 2 |
| 2va7 | C27 | 2.2 |
| 2vij | C44 | 1.6 |
| 2vnn | CM7 | 1.87 |
| 2vnm | CM8 | 1.79 |
| 2ofo | CMZ | 2.25 |
| 2qmf | CS9 | 1.75 |
| 3kmy | D8Y | 1.9 |
| 3hw1 | EV2 | 2.48 |
| 3msk | EV4 | 2 |
| 2zjj | F1J | 2.2 |
| 2zjk | F1K | 3 |
| 3kmx | G00 | 1.7 |
| 2irz | I02 | 1.8 |
| 2oht | IP6 | 2.3 |
| 2qzl | IXS | 1.8 |
| 2ntr | L00 | 1.8 |
| 2hm1 | LIQ | 2.2 |
| 1xs7 | MMI | 2.8 |
| 2oah | QIN | 1.8 |
| 2qmg | SC6 | 1.89 |
| 2qp8 | SC7 | 1.5 |
| 3fkt | SII | 1.9 |

| | | |
|---|---|---|
| 2vie | VG0 | 1.9 |
| 2vj6 | VG5 | 1.8 |
| 2vj7 | VG6 | 1.6 |
| 2vj9 | VG7 | 1.6 |
| 3k5d | XLI | 2.9 |
| 2q11 | XX4 | 2.4 |
| 3lpk | Z76 | 1.93 |
| 2wf1 | ZY1 | 1.6 |
| 2wf3 | ZY3 | 2.08 |

Table B.3.39 PDB structures used to construct the CRANkS grids for run 1 of 50 structures for target BACE1

BACE1, 50 structures, Run 2

| PDB Code | Ligand Residue Code | Resolution / Å |
|---|---|---|
| 3ckr | 9 | 2.7 |
| 3kyr | 38 | 2.6 |
| 3k5c | 0BI | 2.12 |
| 3ivh | 1LI | 1.8 |
| 2zdz | 310 | 2 |
| 3cid | 318 | 1.8 |
| 2b8v | 3BN | 1.8 |
| 2qu3 | 462 | 2 |
| 3in3 | 472 | 2 |
| 2b8l | 5HA | 1.7 |
| 3exo | 5MS | 2.1 |
| 3l3a | 625 | 2.362 |
| 2ph6 | 712 | 2 |
| 3lnk | 74A | 1.8 |
| 3nsh | 957 | 2.2 |
| 2ohs | 9IP | 2.45 |
| 3buf | AEG | 2.3 |
| 2f3e | AXQ | 2.11 |
| 3k5f | AYH | 2.25 |
| 3l5c | BDQ | 1.8 |
| 3l5d | BDV | 1.75 |
| 3l5e | BDW | 1.53 |
| 3l5f | BDX | 1.7 |
| 3k5g | BJC | 2 |
| 2vkm | BSD | 2.05 |
| 2vnn | CM7 | 1.87 |
| 2vnm | CM8 | 1.79 |
| 2qmf | CS9 | 1.75 |
| 3kmy | D8Y | 1.9 |
| 3hvg | EV0 | 2.26 |
| 3msl | EV5 | 2.4 |
| 2zjn | F1N | 2.7 |
| 2iqg | F2I | 1.7 |
| 3kmx | G00 | 1.7 |
| 2va6 | H24 | 2.5 |
| 2qzk | I21 | 1.8 |

| | | |
|---|---|---|
| 2ohu | IP7 | 2.35 |
| 2qzl | IXS | 1.8 |
| 1xs7 | MMI | 2.8 |
| 2oah | QIN | 1.8 |
| 2qmg | SC6 | 1.89 |
| 2viy | VG3 | 1.82 |
| 2viz | VG4 | 1.6 |
| 2vj6 | VG5 | 1.8 |
| 2vj7 | VG6 | 1.6 |
| 3ine | X17 | 1.996 |
| 3k5d | XLI | 2.9 |
| 3lpk | Z76 | 1.93 |
| 3lhg | Z81 | 2.1 |
| 2wez | ZYE | 1.7 |

Table B.3.40 PDB structures used to construct the CRANkS grids for run 2 of 50 structures for target BACE1

BACE1, 50 structures, Run 3

| PDB Code | Ligand Residue Code | Resolution / Å |
|---|---|---|
| 3k5c | 0BI | 2.12 |
| 3ivh | 1LI | 1.8 |
| 2p4j | 23I | 2.5 |
| 2ohl | 2AQ | 2.65 |
| 2b8v | 3BN | 1.8 |
| 2ze1 | 4I1 | 2.2 |
| 3igb | 454 | 2.238 |
| 3in3 | 472 | 2 |
| 3exo | 5MS | 2.1 |
| 2ohp | 6IP | 2.25 |
| 3l38 | 879 | 2.1 |
| 2xfk | AA9 | 1.8 |
| 3buh | AED | 2.3 |
| 3bra | AEF | 2.3 |
| 3dv1 | AR9 | 2.1 |
| 3k5f | AYH | 2.25 |
| 3dv5 | BAV | 2.1 |
| 3l5e | BDW | 1.53 |
| 3l5f | BDX | 1.7 |
| 3in4 | BX2 | 2.3 |
| 2va5 | C8C | 2.75 |
| 2vnm | CM8 | 1.79 |
| 2ofo | CMZ | 2.25 |
| 3l58 | CS5 | 1.8 |
| 2qmf | CS9 | 1.75 |
| 3hvg | EV0 | 2.26 |
| 3hw1 | EV2 | 2.48 |
| 3msk | EV4 | 2 |
| 2zji | F1I | 2.3 |
| 3kmx | G00 | 1.7 |
| 2va6 | H24 | 2.5 |
| 2qzk | I21 | 1.8 |

| | | |
|---|---|---|
| 2ohu | IP7 | 2.35 |
| 2qzl | IXS | 1.8 |
| 1w51 | L01 | 2.55 |
| 2hm1 | LIQ | 2.2 |
| 1xs7 | MMI | 2.8 |
| 2wjo | QUD | 2.5 |
| 2viy | VG3 | 1.82 |
| 3inf | X45 | 1.852 |
| 2xfi | XFI | 1.73 |
| 3k5d | XLI | 2.9 |
| 2q11 | XX4 | 2.4 |
| 3lpi | Z74 | 2.05 |
| 3lpj | Z75 | 1.79 |
| 2g94 | ZPQ | 1.86 |
| 2wfo | ZY0 | 1.6 |
| 2wf1 | ZY1 | 1.6 |
| 2wf3 | ZY3 | 2.08 |
| 2wez | ZYE | 1.7 |

Table B.3.41 PDB structures used to construct the CRANkS grids for run 3 of 50 structures for target BACE1

BACE1, 50 structures, Run 4

| PDB Code | Ligand Residue Code | Resolution / Å |
|---|---|---|
| 3ckp | 12 | 2.3 |
| 3k5c | 0BI | 2.12 |
| 1ym4 | 24M | 2.25 |
| 2zdz | 310 | 2 |
| 2b8v | 3BN | 1.8 |
| 3kno | 3TO | 1.9 |
| 2ze1 | 411 | 2.2 |
| 2b8l | 5HA | 1.7 |
| 3l3a | 625 | 2.362 |
| 2ohp | 6IP | 2.25 |
| 2ohq | 7IP | 2.1 |
| 2ohr | 8IP | 2.25 |
| 2ohs | 9IP | 2.45 |
| 3duy | AFJ | 1.97 |
| 2f3e | AXQ | 2.11 |
| 3hob | B35 | 2.7 |
| 3dv5 | BAV | 2.1 |
| 3l5e | BDW | 1.53 |
| 3l5f | BDX | 1.7 |
| 2va7 | C27 | 2.2 |
| 2vnm | CM8 | 1.79 |
| 2qk5 | CS5 | 2.2 |
| 2qmf | CS9 | 1.75 |
| 3hw1 | EV2 | 2.48 |
| 3msj | EV3 | 1.8 |
| 3msk | EV4 | 2 |
| 3msl | EV5 | 2.4 |

| | | |
|---|---|---|
| 2zji | F1I | 2.3 |
| 2zjl | F1L | 2.1 |
| 2zjm | F1M | 1.9 |
| 3kmx | G00 | 1.7 |
| 2irz | I02 | 1.8 |
| 2ohu | IP7 | 2.35 |
| 2qzl | IXS | 1.8 |
| 1w51 | L01 | 2.55 |
| 1xs7 | MMI | 2.8 |
| 2qmg | SC6 | 1.89 |
| 2viy | VG3 | 1.82 |
| 2viz | VG4 | 1.6 |
| 2xfj | VG5 | 1.8 |
| 2vj7 | VG6 | 1.6 |
| 2vj9 | VG7 | 1.6 |
| 3ine | X17 | 1.996 |
| 3k5d | XLI | 2.9 |
| 3lpi | Z74 | 2.05 |
| 3lpj | Z75 | 1.79 |
| 3lpk | Z76 | 1.93 |
| 2wf1 | ZY1 | 1.6 |
| 2wf4 | ZY4 | 1.8 |
| 2wez | ZYE | 1.7 |

Table B.3.42 PDB structures used to construct the CRANkS grids for run 4 of 50 structures for target BACE1

BACE1, 50 structures , Run 4

| PDB Code | Ligand Residue Code | Resolution / Å |
|---|---|---|
| 1xn2 | 1OL | 1.9 |
| 1ym4 | 24M | 2.25 |
| 3ivi | 2LI | 2.2 |
| 3cib | 314 | 1.72 |
| 3kno | 3TO | 1.9 |
| 2ze1 | 4I1 | 2.2 |
| 2qu3 | 462 | 2 |
| 2ohn | 4FP | 2.15 |
| 3inh | 569 | 1.8 |
| 2b8l | 5HA | 1.7 |
| 3exo | 5MS | 2.1 |
| 2ohp | 6IP | 2.25 |
| 2ohm | 8AP | 2.7 |
| 2ohs | 9IP | 2.45 |
| 2xfk | AA9 | 1.8 |
| 3bra | AEF | 2.3 |
| 3bug | AEH | 2.5 |
| 3dv1 | AR9 | 2.1 |
| 3dv5 | BAV | 2.1 |
| 3l59 | BDJ | 2 |
| 3l5b | BDO | 1.8 |
| 3l5e | BDW | 1.53 |
| 2va7 | C27 | 2.2 |

| | | |
|---|---|---|
| 2vnm | CM8 | 1.79 |
| 2ofo | CMZ | 2.25 |
| 2qmd | CS7 | 1.65 |
| 3kmy | D8Y | 1.9 |
| 3hvg | EV0 | 2.26 |
| 3msj | EV3 | 1.8 |
| 2zjh | F1H | 2.6 |
| 2zji | F1I | 2.3 |
| 2zjl | F1L | 2.1 |
| 3kmx | G00 | 1.7 |
| 2ntr | L00 | 1.8 |
| 2hiz | LIJ | 2.5 |
| 1xs7 | MMI | 2.8 |
| 2qmg | SC6 | 1.89 |
| 2qp8 | SC7 | 1.5 |
| 3fkt | SII | 1.9 |
| 2viy | VG3 | 1.82 |
| 2vj6 | VG5 | 1.8 |
| 2xfj | VG5 | 1.8 |
| 2vj7 | VG6 | 1.6 |
| 3ine | X17 | 1.996 |
| 3k5d | XLI | 2.9 |
| 2q11 | XX4 | 2.4 |
| 3lpi | Z74 | 2.05 |
| 3lpj | Z75 | 1.79 |
| 3lpk | Z76 | 1.93 |
| 2wez | ZYE | 1.7 |

Table B.3.43 PDB structures used to construct the CRANkS grids for run 5 of 50 structures for target BACE1

GRIA2, 5 structures, Run 1

| PDB Code | Ligand Residue Code | Resolution / Å |
|---|---|---|
| 1m5e | AM1 | 1.46 |
| 1mqi | FWD | 1.35 |
| 1nnk | CE2 | 1.85 |
| 3bki | FQX | 1.87 |
| 3ijx | E | 2.881 |

Table B.3.44 PDB structures used to construct the CRANkS grids for run 1 of 5 structures for target GRIA2

GRIA2, 5 structures, Run 2

| PDB Code | Ligand Residue Code | Resolution / Å |
|---|---|---|
| 1lb9 | DNQ | 2.3 |
| 1m5e | AM1 | 1.46 |
| 1not | AT1 | 2.1 |
| 3bki | FQX | 1.87 |
| 3ijx | E | 2.881 |

Table B.3.45 PDB structures used to construct the CRANkS grids for run 2 of 5 structures for target GRIA2

GRIA2, 5 structures, Run 3

| PDB Code | Ligand Residue Code | Resolution / Å |
|---|---|---|
| 1ftk | KAI | 1.6 |
| 1ftl | DNQ | 1.8 |
| 3bki | FQX | 1.87 |
| 3pd8 | HA7 | 2.476 |
| 3ijx | E | 2.881 |

Table B.3.46 PDB structures used to construct the CRANkS grids for run 3 of 5 structures for target GRIA2

GRIA2, 5 structures, Run 4

| PDB Code | Ligand Residue Code | Resolution / Å |
|---|---|---|
| 1ftk | KAI | 1.6 |
| 1m5e | AM1 | 1.46 |
| 1mqh | BWD | 1.8 |
| 1syh | CPW | 1.8 |
| 2cmo | M1L | 2.65 |

Table B.3.47 PDB structures used to construct the CRANkS grids for run 4 of 5 structures for target GRIA2

GRIA2, 5 structures, Run 5

| PDB Code | Ligand Residue Code | Resolution / Å |
|---|---|---|
| 1ftk | KAI | 1.6 |
| 1ftm | AMQ | 1.7 |
| 1lb9 | DNQ | 2.3 |
| 1m5e | AM1 | 1.46 |
| 1mqh | BWD | 1.8 |

Table B.3.48 PDB structures used to construct the CRANkS grids for run 5 of 5 structures for target GRIA2

GRIA2, 10 structures, Run 1

| PDB Code | Ligand Residue Code | Resolution / Å |
|---|---|---|
| 1mm6 | QUS | 2.15 |
| 1mqd | SHI | 1.46 |
| 1mqi | FWD | 1.35 |
| 1not | AT1 | 2.1 |
| 1nnk | CE2 | 1.85 |
| 2aix | U1K | 2.17 |
| 2cmo | M1L | 2.65 |
| 3bki | FQX | 1.87 |
| 3ho6 | VBP | 2.8 |
| 3kgc | ZK1 | 1.55 |

Table B.3.49 PDB structures used to construct the CRANkS grids for run 1 of 10 structures for target GRIA2

GRIA2, 10 structures, Run 2

| PDB Code | Ligand Residue Code | Resolution / Å |
|---|---|---|
| 1ftk | KAI | 1.6 |
| 1ftm | AMQ | 1.7 |
| 1lb9 | DNQ | 2.3 |
| 1mm6 | QUS | 2.15 |
| 1mqj | HWD | 1.65 |
| 1not | AT1 | 2.1 |
| 3b7d | CNI | 2.5 |
| 3bki | FQX | 1.87 |
| 3h03 | UBP | 1.9 |
| 3h06 | VBP | 2.8 |

Table B.3.50 PDB structures used to construct the CRANkS grids for run 2 of 10 structures for target GRIA2

GRIA2, 10 structures, Run 3

| PDB Code | Ligand Residue Code | Resolution / Å |
|---|---|---|
| 1ftl | DNQ | 1.8 |
| 1ftm | AMQ | 1.7 |
| 1m5b | BN1 | 1.85 |
| 1mqd | SHI | 1.46 |
| 1mqj | HWD | 1.65 |
| 2aix | U1K | 2.17 |
| 3b7d | CNI | 2.5 |
| 3bfu | R2P | 1.95 |
| 3h06 | VBP | 2.8 |
| 3pd8 | HA7 | 2.476 |

Table B.3.51 PDB structures used to construct the CRANkS grids for run 3 of 10 structures for target GRIA2

GRIA2, 10 structures, Run 4

| PDB Code | Ligand Residue Code | Resolution / Å |
|---|---|---|
| 1ftk | KAI | 1.6 |
| 1ftl | DNQ | 1.8 |
| 1m5e | AM1 | 1.46 |
| 1mm6 | QUS | 2.15 |
| 1mqh | BWD | 1.8 |
| 1syh | CPW | 1.8 |
| 2aix | U1K | 2.17 |
| 2cmo | M1L | 2.65 |
| 3h06 | VBP | 2.8 |
| 3pd9 | HA5 | 2.1 |

Table B.3.52 PDB structures used to construct the CRANkS grids for run 4 of 10 structures for target GRIA2

GRIA2, 10 structures, Run 5

| PDB Code | Ligand Residue Code | Resolution / Å |
|---|---|---|
| 1ftl | DNQ | 1.8 |
| 1ftm | AMQ | 1.7 |
| 1m5b | BN1 | 1.85 |
| 1m5e | AM1 | 1.46 |
| 1mm6 | QUS | 2.15 |
| 1mqi | FWD | 1.35 |
| 1syh | CPW | 1.8 |
| 3bki | FQX | 1.87 |
| 3h03 | UBP | 1.9 |
| 3kgc | ZK1 | 1.55 |

Table B.3.53 PDB structures used to construct the CRANkS grids for run 5 of 10 structures for target GRIA2

GRIA2, 25 structures, Run 1

| PDB Code | Ligand Residue Code | Resolution / Å |
|---|---|---|
| 1ftj | E | 1.9 |
| 1ftl | DNQ | 1.8 |
| 1ftm | AMQ | 1.7 |
| 1m5b | BN1 | 1.85 |
| 1m5e | AM1 | 1.46 |
| 1mm6 | QUS | 2.15 |
| 1mqd | SHI | 1.46 |
| 1mqh | BWD | 1.8 |
| 1mqi | FWD | 1.35 |
| 1mqj | HWD | 1.65 |
| 1not | AT1 | 2.1 |
| 1nnk | CE2 | 1.85 |
| 1syh | CPW | 1.8 |
| 2aix | U1K | 2.17 |
| 2cmo | M1L | 2.65 |
| 3bft | S2P | 2.27 |
| 3bfu | R2P | 1.95 |
| 3bki | FQX | 1.87 |
| 3h03 | UBP | 1.9 |
| 3h06 | VBP | 2.8 |
| 3ijx | E | 2.881 |
| 3kg2 | ZK1 | 3.6 |
| 3kgc | ZK1 | 1.55 |

| | | |
|---|---|---|
| 3pd8 | HA7 | 2.476 |
| 3pd9 | HA5 | 2.1 |

Table B.3.54 PDB structures used to construct the CRANkS grids for run 1 of 25 structures for target GRIA2

GRIA2, 25 structures, Run 2

| PDB Code | Ligand Residue Code | Resolution / Å |
|---|---|---|
| 1ftj | E | 1.9 |
| 1ftl | DNQ | 1.8 |
| 1ftm | AMQ | 1.7 |
| 1lb9 | DNQ | 2.3 |
| 1m5b | BN1 | 1.85 |
| 1m5c | BRH | 1.65 |
| 1m5e | AM1 | 1.46 |
| 1mm6 | QUS | 2.15 |
| 1mqd | SHI | 1.46 |
| 1mqh | BWD | 1.8 |
| 1mqi | FWD | 1.35 |
| 1mqj | HWD | 1.65 |
| 1not | AT1 | 2.1 |
| 1syh | CPW | 1.8 |
| 2aix | U1K | 2.17 |
| 3b7d | CNI | 2.5 |
| 3bft | S2P | 2.27 |
| 3bfu | R2P | 1.95 |
| 3bki | FQX | 1.87 |
| 3h03 | UBP | 1.9 |
| 3h06 | VBP | 2.8 |
| 3ijx | E | 2.881 |
| 3kgc | ZK1 | 1.55 |
| 3pd8 | HA7 | 2.476 |
| 3pd9 | HA5 | 2.1 |

Table B.3.55 PDB structures used to construct the CRANkS grids for run 2 of 25 structures for target GRIA2

GRIA2, 25 structures, Run 3

| PDB Code | Ligand Residue Code | Resolution / Å |
|---|---|---|
| 1ftj | E | 1.9 |
| 1ftk | KAI | 1.6 |
| 1ftl | DNQ | 1.8 |
| 1lb9 | DNQ | 2.3 |
| 1m5b | BN1 | 1.85 |
| 1m5e | AM1 | 1.46 |
| 1mm6 | QUS | 2.15 |
| 1mqd | SHI | 1.46 |
| 1mqg | IWD | 2.15 |
| 1mqh | BWD | 1.8 |

| | | |
|---|---|---|
| 1mqi | FWD | 1.35 |
| 1mqj | HWD | 1.65 |
| 1nnk | CE2 | 1.85 |
| 1syh | CPW | 1.8 |
| 2aix | U1K | 2.17 |
| 2cmo | M1L | 2.65 |
| 3bft | S2P | 2.27 |
| 3bfu | R2P | 1.95 |
| 3bki | FQX | 1.87 |
| 3h03 | UBP | 1.9 |
| 3ijx | E | 2.881 |
| 3kg2 | ZK1 | 3.6 |
| 3kgc | ZK1 | 1.55 |
| 3pd8 | HA7 | 2.476 |
| 3pd9 | HA5 | 2.1 |

Table B.3.56 PDB structures used to construct the CRANkS grids for run 3 of 25 structures for target GRIA2

GRIA2, 25 structures, Run 4

| PDB Code | Ligand Residue Code | Resolution / Å |
|---|---|---|
| 1ftj | E | 1.9 |
| 1ftk | KAI | 1.6 |
| 1ftl | DNQ | 1.8 |
| 1lb9 | DNQ | 2.3 |
| 1m5c | BRH | 1.65 |
| 1m5e | AM1 | 1.46 |
| 1mqd | SHI | 1.46 |
| 1mqg | IWD | 2.15 |
| 1mqh | BWD | 1.8 |
| 1mqj | HWD | 1.65 |
| 1not | AT1 | 2.1 |
| 1nnk | CE2 | 1.85 |
| 1syh | CPW | 1.8 |
| 2aix | U1K | 2.17 |
| 2cmo | M1L | 2.65 |
| 3b7d | CNI | 2.5 |
| 3bft | S2P | 2.27 |
| 3bfu | R2P | 1.95 |
| 3bki | FQX | 1.87 |
| 3h06 | VBP | 2.8 |
| 3ijx | E | 2.881 |
| 3kg2 | ZK1 | 3.6 |
| 3kgc | ZK1 | 1.55 |
| 3pd8 | HA7 | 2.476 |

| | | |
|---|---|---|
| 3pd9 | HA5 | 2.1 |

Table B.3.57 PDB structures used to construct the CRANkS grids for run 4 of 25 structures for target GRIA2

GRIA2, 25 structures, Run 5

| PDB Code | Ligand Residue Code | Resolution / Å |
|---|---|---|
| 1ftk | KAI | 1.6 |
| 1ftl | DNQ | 1.8 |
| 1ftm | AMQ | 1.7 |
| 1lb9 | DNQ | 2.3 |
| 1m5b | BN1 | 1.85 |
| 1m5c | BRH | 1.65 |
| 1m5e | AM1 | 1.46 |
| 1mm6 | QUS | 2.15 |
| 1mqd | SHI | 1.46 |
| 1mqg | IWD | 2.15 |
| 1mqh | BWD | 1.8 |
| 1mqi | FWD | 1.35 |
| 1not | AT1 | 2.1 |
| 1nnk | CE2 | 1.85 |
| 1syh | CPW | 1.8 |
| 2aix | U1K | 2.17 |
| 2cmo | M1L | 2.65 |
| 3bft | S2P | 2.27 |
| 3bfu | R2P | 1.95 |
| 3bki | FQX | 1.87 |
| 3h06 | VBP | 2.8 |
| 3ijx | E | 2.881 |
| 3kg2 | ZK1 | 3.6 |
| 3kgc | ZK1 | 1.55 |
| 3pd8 | HA7 | 2.476 |

Table B.3.58 PDB structures used to construct the CRANkS grids for run 5 of 25 structures for target GRIA2

HSP90A, 5 structures, Run 1

| PDB Code | Ligand Residue Code | Resolution / Å |
|---|---|---|
| 1yet | GDM | 1.9 |
| 2fwy | H64 | 2.1 |
| 2xk2 | ADP | 1.95 |
| 3bmy | CX2 | 1.6 |
| 3k98 | 1RC | 2.4 |

Table B.3.59 PDB structures used to construct the CRANkS grids for run 1 of 5 structures for target HSP90A

HSP90A, 5 structures, Run 2

| PDB Code | Ligand Residue Code | Resolution / Å |
|---|---|---|
| 1uy6 | PU3 | 1.9 |
| 1uy8 | PU5 | 1.98 |
| 2h55 | DZ8 | 2 |
| 2wi6 | ZZ6 | 2.18 |
| 2xk2 | ADP | 1.95 |

Table B.3.60 PDB structures used to construct the CRANkS grids for run 2 of 5 structures for target HSP90A

HSP90A, 5 structures, Run 3

| PDB Code | Ligand Residue Code | Resolution / Å |
|---|---|---|
| 2bto | CT5 | 1.9 |
| 2qf6 | A56 | 3.1 |
| 3ekr | PY9 | 2 |
| 3ft5 | MO8 | 1.9 |
| 3ft8 | MOJ | 2 |

Table B.3.61 PDB structures used to construct the CRANkS grids for run 3 of 5 structures for target HSP90A

HSP90A, 5 structures, Run 4

| PDB Code | Ligand Residue Code | Resolution / Å |
|---|---|---|
| 1uye | PU9 | 2 |
| 1uyk | PUX | 2.2 |
| 2ccu | 2D9 | 2.7 |
| 2vcj | 2EQ | 2.5 |
| 3hhu | 819 | 1.59 |

Table B.3.62 PDB structures used to construct the CRANkS grids for run 4 of 5 structures for target HSP90A

HSP90A, 5 structures, Run 5

| PDB Code | Ligand Residue Code | Resolution / Å |
|---|---|---|
| 1uyd | PU8 | 2.2 |
| 2bto | CT5 | 1.9 |
| 2fwy | H64 | 2.1 |
| 2wi6 | ZZ6 | 2.18 |
| 3hz5 | Z64 | 1.9 |

Table B.3.63 PDB structures used to construct the CRANkS grids for run 5 of 5 structures for target HSP90A

HSP90A, 10 structures, Run 1

| PDB Code | Ligand Residue Code | Resolution / Å |
|---|---|---|
| 1uy6 | PU3 | 1.9 |

| | | |
|---|---|---|
| 1uy9 | PU6 | 2 |
| 1yc1 | 4BC | 1.7 |
| 2qf6 | A56 | 3.1 |
| 2wi5 | ZZ5 | 2.1 |
| 2wi6 | ZZ6 | 2.18 |
| 2wi7 | 2KL | 2.5 |
| 2xdk | XDK | 1.97 |
| 2xdl | 2DL | 1.98 |
| 3k99 | PFT | 2.1 |

Table B.3.64 PDB structures used to construct the CRANkS grids for run 1 of 10 structures for target HSP90A

HSP90A, 10 structures, Run 2

| PDB Code | Ligand Residue Code | Resolution / Å |
|---|---|---|
| 1uyk | PUX | 2.2 |
| 1yc1 | 4BC | 1.7 |
| 2bsm | BSM | 2.05 |
| 2ccs | 4BH | 1.79 |
| 2ccu | 2D9 | 2.7 |
| 2qg2 | A91 | 1.8 |
| 2wi2 | ZZ3 | 2.09 |
| 2xht | CoY | 2.27 |
| 3bmy | CX2 | 1.6 |

Table B.3.65 PDB structures used to construct the CRANkS grids for run 2 of 10 structures for target HSP90A

HSP90A, 10 structures, Run 3

| PDB Code | Ligand Residue Code | Resolution / Å |
|---|---|---|
| 1uye | PU9 | 2 |
| 1uyi | PUZ | 2.2 |
| 1yc1 | 4BC | 1.7 |
| 1yc4 | 43P | 1.81 |
| 2bto | CT5 | 1.9 |
| 2wi3 | ZZ3 | 1.9 |
| 2xab | VHD | 1.9 |
| 3hek | BDo | 1.95 |
| 3inx | JZC | 1.75 |

Table B.3.66 PDB structures used to construct the CRANkS grids for run 3 of 10 structures for target HSP90A

HSP90A, 10 structures, Run 4

| PDB Code | Ligand Residue Code | Resolution / Å |
|---|---|---|
| 1uy7 | PU4 | 1.9 |
| 1uy8 | PU5 | 1.98 |
| 1uyd | PU8 | 2.2 |

| | | |
|---|---|---|
| 1uye | PU9 | 2 |
| 1uyi | PUZ | 2.2 |
| 2byi | 2DD | 1.6 |
| 2vcj | 2EQ | 2.5 |
| 2wi2 | ZZ3 | 2.09 |
| 2xdk | XDK | 1.97 |
| 3hz1 | 37D | 2.3 |

Table B.3.67 PDB structures used to construct the CRANkS grids for run 4 of 10 structures for target HSP90A

HSP90A, 10 structures, Run 5

| PDB Code | Ligand Residue Code | Resolution / Å |
|---|---|---|
| 1uye | PU9 | 2 |
| 1yc3 | 4BC | 2.12 |
| 2byi | 2DD | 1.6 |
| 2qfo | A13 | 1.68 |
| 2qfo | A51 | 1.68 |
| 2qg2 | A91 | 1.8 |
| 2wi2 | ZZ3 | 2.09 |
| 2wi4 | ZZ4 | 2.4 |
| 3hz1 | 37D | 2.3 |
| 3hz5 | Z64 | 1.9 |
| 3inx | JZC | 1.75 |

Table B.3.68 PDB structures used to construct the CRANkS grids for run 5 of 10 structures for target HSP90A

HSP90A, 25 structures, Run 1

| PDB Code | Ligand Residue Code | Resolution / Å |
|---|---|---|
| 1uye | PU9 | 2 |
| 2byh | 2D7 | 1.9 |
| 2byi | 2DD | 1.6 |
| 2bz5 | AB4 | 1.9 |
| 2cct | 2E0 | 2.3 |
| 2qfo | A13 | 1.68 |
| 2qfo | A51 | 1.68 |
| 2qg0 | A94 | 1.85 |
| 2vci | 2GJ | 2 |
| 2wi1 | ZZ2 | 2.3 |
| 2wi2 | ZZ3 | 2.09 |
| 2wi7 | 2KL | 2.5 |
| 2xdl | 2DL | 1.98 |
| 2xht | CoY | 2.27 |
| 2xjx | XJX | 1.66 |
| 3bm9 | BX2 | 1.6 |
| 3dob | SNX | 1.74 |
| 3eko | PYU | 1.55 |
| 3ekr | PY9 | 2 |

| | | |
|---|---|---|
| 3hek | BDo | 1.95 |
| 3hhu | 819 | 1.59 |
| 3hz1 | 37D | 2.3 |
| 3hz5 | Z64 | 1.9 |
| 3k98 | 1RC | 2.4 |
| 3k99 | PFT | 2.1 |

Table B.3.69 PDB structures used to construct the CRANkS grids for run 1 of 25 structures for target HSP90A

HSP90A, 25 structures, Run 2

| PDB Code | Ligand Residue Code | Resolution / Å |
|---|---|---|
| 1osf | KOS | 1.75 |
| 1uy6 | PU3 | 1.9 |
| 1uy7 | PU4 | 1.9 |
| 1uye | PU9 | 2 |
| 1uyh | PUo | 2.2 |
| 1yet | GDM | 1.9 |
| 2bsm | BSM | 2.05 |
| 2byh | 2D7 | 1.9 |
| 2ccs | 4BH | 1.79 |
| 2jjc | LGA | 1.95 |
| 2qfo | A13 | 1.68 |
| 2qfo | A51 | 1.68 |
| 2qgo | A94 | 1.85 |
| 2qg2 | A91 | 1.8 |
| 2uwd | 2GG | 1.9 |
| 2wi1 | ZZ2 | 2.3 |
| 2wi4 | ZZ4 | 2.4 |
| 2wi5 | ZZ5 | 2.1 |
| 2xdx | WOE | 2.42 |
| 2xjj | L81 | 1.9 |
| 3bm9 | BX2 | 1.6 |
| 3dob | SNX | 1.74 |
| 3ft8 | MOJ | 2 |
| 3hek | BDo | 1.95 |
| 3hz1 | 37D | 2.3 |
| 3inx | JZC | 1.75 |

Table B.3.70 PDB structures used to construct the CRANkS grids for run 2 of 25 structures for target HSP90A

HSP90A, 25 structures, Run 3

| PDB Code | Ligand Residue Code | Resolution / Å |
|---|---|---|
| 1uy7 | PU4 | 1.9 |
| 1uy9 | PU6 | 2 |
| 1uyd | PU8 | 2.2 |

| | | |
|---|---|---|
| 1uyk | PUX | 2.2 |
| 2byi | 2DD | 1.6 |
| 2cct | 2E1 | 2.3 |
| 2ccu | 2D9 | 2.7 |
| 2jjc | LGA | 1.95 |
| 2qf6 | A56 | 3.1 |
| 2qg0 | A94 | 1.85 |
| 2qg2 | A91 | 1.8 |
| 2uwd | 2GG | 1.9 |
| 2wi3 | ZZ3 | 1.9 |
| 2wi4 | ZZ4 | 2.4 |
| 2wi6 | ZZ6 | 2.18 |
| 2wi7 | 2KL | 2.5 |
| 2xdk | XDK | 1.97 |
| 2xhr | CoP | 2.2 |
| 2xjj | L81 | 1.9 |
| 2xjx | XJX | 1.66 |
| 3ft8 | MOJ | 2 |
| 3hyz | 42C | 2.3 |
| 3inx | JZC | 1.75 |
| 3k97 | 4CD | 1.95 |
| 3mnr | SD1 | 1.9 |

Table B.3.71 PDB structures used to construct the CRANkS grids for run 3 of 25 structures for target HSP90A

HSP90A, 25 structures, Run 4

| PDB Code | Ligand Residue Code | Resolution / Å |
|---|---|---|
| 1uy6 | PU3 | 1.9 |
| 1uy7 | PU4 | 1.9 |
| 1uy9 | PU6 | 2 |
| 1uyc | PU7 | 2 |
| 1uyi | PUZ | 2.2 |
| 2bto | CT5 | 1.9 |
| 2byh | 2D7 | 1.9 |
| 2cct | 2E1 | 2.3 |
| 2h55 | DZ8 | 2 |
| 2vci | 2GJ | 2 |
| 2wi3 | ZZ3 | 1.9 |
| 2wi4 | ZZ4 | 2.4 |
| 2xab | VHD | 1.9 |
| 2xdk | XDK | 1.97 |
| 2xds | MTo | 1.97 |
| 2xdu | LGA | 1.74 |
| 2xhr | CoP | 2.2 |
| 2xjg | XJG | 2.25 |
| 2xjj | L81 | 1.9 |
| 2xk2 | ADP | 1.95 |
| 3bm9 | BX2 | 1.6 |
| 3ft8 | MOJ | 2 |
| 3hhu | 819 | 1.59 |

| | | |
|---|---|---|
| 3hyy | 37D | 1.9 |
| 3mnr | SD1 | 1.9 |

Table B.3.72 PDB structures used to construct the CRANkS grids for run 4 of 25 structures for target HSP90A

HSP90A, 25 structures, Run 5

| PDB Code | Ligand Residue Code | Resolution / Å |
|---|---|---|
| 1uyc | PU7 | 2 |
| 1uyd | PU8 | 2.2 |
| 1uye | PU9 | 2 |
| 1uyf | PU1 | 2 |
| 1uyi | PUZ | 2.2 |
| 1yc1 | 4BC | 1.7 |
| 1yc3 | 4BC | 2.12 |
| 1yet | GDM | 1.9 |
| 2bsm | BSM | 2.05 |
| 2bto | CT5 | 1.9 |
| 2bz5 | AB4 | 1.9 |
| 2cct | 2E1 | 2.3 |
| 2qf6 | A56 | 3.1 |
| 2wi4 | ZZ4 | 2.4 |
| 2xdk | XDK | 1.97 |
| 2xds | MTo | 1.97 |
| 2xjx | XJX | 1.66 |
| 3bm9 | BX2 | 1.6 |
| 3eko | PYU | 1.55 |
| 3ekr | PY9 | 2 |
| 3ft5 | MO8 | 1.9 |
| 3hek | BDo | 1.95 |
| 3hyy | 37D | 1.9 |
| 3hz5 | Z64 | 1.9 |
| 3inw | JZB | 1.95 |

Table B.3.73 PDB structures used to construct the CRANkS grids for run 5 of 25 structures for target HSP90A

HSP90A, 50 structures, Run 1

| PDB Code | Ligand Residue Code | Resolution / Å |
|---|---|---|
| 1osf | KOS | 1.75 |
| 1uy6 | PU3 | 1.9 |
| 1uyd | PU8 | 2.2 |
| 1uye | PU9 | 2 |
| 1uyf | PU1 | 2 |
| 1yc3 | 4BC | 2.12 |
| 2bto | CT5 | 1.9 |
| 2byh | 2D7 | 1.9 |

| | | |
|---|---|---|
| 2byi | 2DD | 1.6 |
| 2bz5 | AB4 | 1.9 |
| 2ccs | 4BH | 1.79 |
| 2cct | 2E1 | 2.3 |
| 2ccu | 2D9 | 2.7 |
| 2cdd | CT5 | -- |
| 2jjc | LGA | 1.95 |
| 2qf6 | A56 | 3.1 |
| 2qfo | A13 | 1.68 |
| 2qfo | A51 | 1.68 |
| 2qg0 | A94 | 1.85 |
| 2qg2 | A91 | 1.8 |
| 2vci | 2GJ | 2 |
| 2wi1 | ZZ2 | 2.3 |
| 2wi2 | ZZ3 | 2.09 |
| 2wi4 | ZZ4 | 2.4 |
| 2wi5 | ZZ5 | 2.1 |
| 2wi6 | ZZ6 | 2.18 |
| 2wi7 | 2KL | 2.5 |
| 2xab | VHD | 1.9 |
| 2xdk | XDK | 1.97 |
| 2xdl | 2DL | 1.98 |
| 2xdx | WOE | 2.42 |
| 2xhr | CoP | 2.2 |
| 2xht | CoY | 2.27 |
| 2xjg | XJG | 2.25 |
| 2xjj | L81 | 1.9 |
| 2xjx | XJX | 1.66 |
| 3bm9 | BX2 | 1.6 |
| 3dob | SNX | 1.74 |
| 3eko | PYU | 1.55 |
| 3ekr | PY9 | 2 |
| 3ft5 | MO8 | 1.9 |
| 3ft8 | MOJ | 2 |
| 3hek | BD0 | 1.95 |
| 3hhu | 819 | 1.59 |
| 3hyy | 37D | 1.9 |
| 3hz1 | 37D | 2.3 |
| 3hz5 | Z64 | 1.9 |
| 3inw | JZB | 1.95 |
| 3k98 | 1RC | 2.4 |
| 3k99 | PFT | 2.1 |

Table B.3.74 PDB structures used to construct the CRANkS grids for run 1 of 50 structures for target HSP90A

HSP90A, 50 structures, Run 2

| PDB Code | Ligand Residue Code | Resolution / Å |
|---|---|---|
| 1osf | KOS | 1.75 |
| 1uy6 | PU3 | 1.9 |

| | | |
|---|---|---|
| 1uy7 | PU4 | 1.9 |
| 1uy8 | PU5 | 1.98 |
| 1uye | PU9 | 2 |
| 1uyg | PU2 | 2 |
| 1uyh | PU0 | 2.2 |
| 1yc1 | 4BC | 1.7 |
| 1yc3 | 4BC | 2.12 |
| 1yet | GDM | 1.9 |
| 2bsm | BSM | 2.05 |
| 2bt0 | CT5 | 1.9 |
| 2byh | 2D7 | 1.9 |
| 2ccs | 4BH | 1.79 |
| 2cct | 2E1 | 2.3 |
| 2cdd | CT5 | -- |
| 2h55 | DZ8 | 2 |
| 2jjc | LGA | 1.95 |
| 2qf6 | A56 | 3.1 |
| 2qfo | A13 | 1.68 |
| 2qfo | A51 | 1.68 |
| 2qg0 | A94 | 1.85 |
| 2qg2 | A91 | 1.8 |
| 2uwd | 2GG | 1.9 |
| 2vcj | 2EQ | 2.5 |
| 2wi1 | ZZ2 | 2.3 |
| 2wi2 | ZZ3 | 2.09 |
| 2wi3 | ZZ3 | 1.9 |
| 2wi4 | ZZ4 | 2.4 |
| 2wi5 | ZZ5 | 2.1 |
| 2xab | VHD | 1.9 |
| 2xdu | LGA | 1.74 |
| 2xdx | WOE | 2.42 |
| 2xhr | CoP | 2.2 |
| 2xjg | XJG | 2.25 |
| 2xjj | L81 | 1.9 |
| 3bm9 | BX2 | 1.6 |
| 3bmy | CX2 | 1.6 |
| 3dob | SNX | 1.74 |
| 3eko | PYU | 1.55 |
| 3ft5 | MO8 | 1.9 |
| 3ft8 | MOJ | 2 |
| 3hek | BD0 | 1.95 |
| 3hhu | 819 | 1.59 |
| 3hyy | 37D | 1.9 |
| 3hz1 | 37D | 2.3 |
| 3inx | JZC | 1.75 |
| 3k97 | 4CD | 1.95 |
| 3k99 | PFT | 2.1 |
| 3mnr | SD1 | 1.9 |

Table B.3.75 PDB structures used to construct the CRANkS grids for run 2 of 50 structures for target HSP90A

HSP90A, 50 structures, Run 3

| PDB Code | Ligand Residue Code | Resolution / Å |
|---|---|---|
| 1uy6 | PU3 | 1.9 |
| 1uy7 | PU4 | 1.9 |
| 1uy9 | PU6 | 2 |
| 1uyd | PU8 | 2.2 |
| 1uyh | PU0 | 2.2 |
| 1uyk | PUX | 2.2 |
| 1yc3 | 4BC | 2.12 |
| 1yet | GDM | 1.9 |
| 2bsm | BSM | 2.05 |
| 2bto | CT5 | 1.9 |
| 2byi | 2DD | 1.6 |
| 2cct | 2E1 | 2.3 |
| 2ccu | 2D9 | 2.7 |
| 2cdd | CT5 | NaN |
| 2fwz | H71 | 2.1 |
| 2jjc | LGA | 1.95 |
| 2qf6 | A56 | 3.1 |
| 2qgo | A94 | 1.85 |
| 2qg2 | A91 | 1.8 |
| 2uwd | 2GG | 1.9 |
| 2vcj | 2EQ | 2.5 |
| 2wi1 | ZZ2 | 2.3 |
| 2wi2 | ZZ3 | 2.09 |
| 2wi3 | ZZ3 | 1.9 |
| 2wi4 | ZZ4 | 2.4 |
| 2wi6 | ZZ6 | 2.18 |
| 2wi7 | 2KL | 2.5 |
| 2xab | VHD | 1.9 |
| 2xdk | XDK | 1.97 |
| 2xdl | 2DL | 1.98 |
| 2xdx | WOE | 2.42 |
| 2xhr | CoP | 2.2 |
| 2xhx | T5M | 2.801 |
| 2xjg | XJG | 2.25 |
| 2xjj | L81 | 1.9 |
| 2xjx | XJX | 1.66 |
| 3bm9 | BX2 | 1.6 |
| 3dob | SNX | 1.74 |
| 3eko | PYU | 1.55 |
| 3ft8 | MOJ | 2 |
| 3hyy | 37D | 1.9 |
| 3hyz | 42C | 2.3 |
| 3hz1 | 37D | 2.3 |
| 3hz5 | Z64 | 1.9 |
| 3inw | JZB | 1.95 |
| 3inx | JZC | 1.75 |

| | | |
|---|---|---|
| 3k97 | 4CD | 1.95 |
| 3k98 | 1RC | 2.4 |
| 3k99 | PFT | 2.1 |
| 3mnr | SD1 | 1.9 |

Table B.3.76 PDB structures used to construct the CRANkS grids for run 3 of 50 structures for target HSP90A

HSP90A, 50 structures, Run 4

| PDB Code | Ligand Residue Code | Resolution / Å |
|---|---|---|
| 1uy6 | PU3 | 1.9 |
| 1uy7 | PU4 | 1.9 |
| 1uy9 | PU6 | 2 |
| 1uyc | PU7 | 2 |
| 1uyd | PU8 | 2.2 |
| 1uye | PU9 | 2 |
| 1uyh | PU0 | 2.2 |
| 1uyi | PUZ | 2.2 |
| 1uyk | PUX | 2.2 |
| 1yet | GDM | 1.9 |
| 2bto | CT5 | 1.9 |
| 2byh | 2D7 | 1.9 |
| 2bz5 | AB4 | 1.9 |
| 2ccs | 4BH | 1.79 |
| 2cct | 2E1 | 2.3 |
| 2fwz | H71 | 2.1 |
| 2h55 | DZ8 | 2 |
| 2jjc | LGA | 1.95 |
| 2qf6 | A56 | 3.1 |
| 2qfo | A13 | 1.68 |
| 2qfo | A51 | 1.68 |
| 2qg0 | A94 | 1.85 |
| 2qg2 | A91 | 1.8 |
| 2vci | 2GJ | 2 |
| 2wi2 | ZZ3 | 2.09 |
| 2wi3 | ZZ3 | 1.9 |
| 2wi4 | ZZ4 | 2.4 |
| 2wi6 | ZZ6 | 2.18 |
| 2wi7 | 2KL | 2.5 |
| 2xab | VHD | 1.9 |
| 2xdk | XDK | 1.97 |
| 2xds | MTo | 1.97 |
| 2xdu | LGA | 1.74 |
| 2xhr | CoP | 2.2 |
| 2xht | CoY | 2.27 |
| 2xjg | XJG | 2.25 |
| 2xjj | L81 | 1.9 |
| 2xk2 | ADP | 1.95 |
| 3bm9 | BX2 | 1.6 |
| 3bmy | CX2 | 1.6 |
| 3eko | PYU | 1.55 |

| | | |
|---|---|---|
| 3ft8 | MOJ | 2 |
| 3hek | BDo | 1.95 |
| 3hhu | 819 | 1.59 |
| 3hyy | 37D | 1.9 |
| 3hz1 | 37D | 2.3 |
| 3hz5 | Z64 | 1.9 |
| 3inw | JZB | 1.95 |
| 3inx | JZC | 1.75 |
| 3k98 | 1RC | 2.4 |
| 3mnr | SD1 | 1.9 |

Table B.3.77 PDB structures used to construct the CRANkS grids for run 4 of 50 structures for target HSP90A

HSP90A, 50 structures, Run 5

| PDB Code | Ligand Residue Code | Resolution / Å |
|---|---|---|
| 1uy8 | PU5 | 1.98 |
| 1uy9 | PU6 | 2 |
| 1uyc | PU7 | 2 |
| 1uyd | PU8 | 2.2 |
| 1uye | PU9 | 2 |
| 1uyf | PU1 | 2 |
| 1uyi | PUZ | 2.2 |
| 1uyk | PUX | 2.2 |
| 1yc1 | 4BC | 1.7 |
| 1yc3 | 4BC | 2.12 |
| 1yc4 | 43P | 1.81 |
| 1yet | GDM | 1.9 |
| 2bsm | BSM | 2.05 |
| 2bto | CT5 | 1.9 |
| 2bz5 | AB4 | 1.9 |
| 2cct | 2E1 | 2.3 |
| 2ccu | 2D9 | 2.7 |
| 2cdd | CT5 | NaN |
| 2fwy | H64 | 2.1 |
| 2fwz | H71 | 2.1 |
| 2qf6 | A56 | 3.1 |
| 2qfo | A13 | 1.68 |
| 2qfo | A51 | 1.68 |
| 2qgo | A94 | 1.85 |
| 2uwd | 2GG | 1.9 |
| 2vci | 2GJ | 2 |
| 2wi1 | ZZ2 | 2.3 |
| 2wi2 | ZZ3 | 2.09 |
| 2wi4 | ZZ4 | 2.4 |
| 2wi5 | ZZ5 | 2.1 |
| 2wi6 | ZZ6 | 2.18 |
| 2wi7 | 2KL | 2.5 |
| 2xdk | XDK | 1.97 |

| | | |
|---|---|---|
| 2xdl | 2DL | 1.98 |
| 2xds | MTo | 1.97 |
| 2xhr | CoP | 2.2 |
| 2xhx | T5M | 2.801 |
| 2xjx | XJX | 1.66 |
| 2xk2 | ADP | 1.95 |
| 3bm9 | BX2 | 1.6 |
| 3eko | PYU | 1.55 |
| 3ekr | PY9 | 2 |
| 3ft5 | MO8 | 1.9 |
| 3hek | BDo | 1.95 |
| 3hhu | 819 | 1.59 |
| 3hyy | 37D | 1.9 |
| 3hyz | 42C | 2.3 |
| 3hz1 | 37D | 2.3 |
| 3hz5 | Z64 | 1.9 |
| 3inw | JZB | 1.95 |
| 3mnr | SD1 | 1.9 |

Table B.3.78 PDB structures used to construct the CRANkS grids for run 5 of 50 structures for target HSP90A

PYGM, 5 structures, Run 1

| PDB Code | Ligand Residue Code | Resolution / Å |
|---|---|---|
| 2f3q | 6GP | 1.96 |
| 1c50 | CHI | 2.3 |
| 3l7d | DK5 | 2 |
| 2fet | H53 | 2.03 |
| 1ww2 | NBG | 1.9 |

Table B.3.79 PDB structures used to construct the CRANkS grids for run 1 of 5 structures for target PYGM

PYGM, 5 structures, Run 2

| PDB Code | Ligand Residue Code | Resolution / Å |
|---|---|---|
| 2g9q | 1AB | 2.5 |
| 2f3q | 6GP | 1.96 |
| 3l7c | DK4 | 1.93 |
| 1z62 | IAA | 1.9 |
| 1ww2 | NBG | 1.9 |

Table B.3.80 PDB structures used to construct the CRANkS grids for run 2 of 5 structures for target PYGM

PYGM, 5 structures, Run 3

| PDB Code | Ligand Residue Code | Resolution / Å |
|---|---|---|
| 3mrt | 12E | 1.98 |
| 3mrx | 17S | 1.95 |
| 2qrm | M09 | 1.9 |
| 3g2i | RUG | 2 |

| | | |
|---|---|---|
| 2qrp | So6 | 1.86 |

Table B.3.81 PDB structures used to construct the CRANkS grids for run 3 of 5 structures for target PYGM

PYGM, 5 structures, Run 4

| PDB Code | Ligand Residue Code | Resolution / Å |
|---|---|---|
| 2f3p | 4GP | 1.94 |
| 3bd8 | C3B | 2.1 |
| 3l7b | DKZ | 2 |
| 2g9r | G27 | 2.07 |
| 4gpb | GFP | 2.3 |

Table B.3.82 PDB structures used to construct the CRANkS grids for run 4 of 5 structures for target PYGM

PYGM, 5 structures, Run 5

| PDB Code | Ligand Residue Code | Resolution / Å |
|---|---|---|
| 3mrv | 16F | 1.94 |
| 3mt8 | 17T | 2 |
| 3mqf | 20X | 1.951 |
| 2f3p | 4GP | 1.94 |
| 1p4j | CBF | 2 |

Table B.3.83 PDB structures used to construct the CRANkS grids for run 5 of 5 structures for target PYGM

PYGM, 10 structures, Run 1

| PDB Code | Ligand Residue Code | Resolution / Å |
|---|---|---|
| 2f3q | 6GP | 1.96 |
| 1z8d | ADE | 2.3 |
| 1c8l | CFF | 2.3 |
| 1c50 | CHI | 2.3 |
| 1fu8 | CR6 | 2.35 |
| 3l7d | DK5 | 2 |
| 2qln | F59 | 2.15 |
| 2fet | H53 | 2.03 |
| 3g2h | KOT | 2.03 |
| 1ww2 | NBG | 1.9 |

Table B.3.84 PDB structures used to construct the CRANkS grids for run 1 of 10 structures for target PYGM

PYGM, 10 structures, Run 2

| PDB Code | Ligand Residue Code | Resolution / Å |
|---|---|---|
| 3cuw | 445 | 2 |
| 3mrx | 17S | 1.95 |
| 2g9q | 1AB | 2.5 |
| 2f3q | 6GP | 1.96 |
| 3ebp | CPB | 2 |
| 3l7c | DK4 | 1.93 |
| 1ftw | GL5 | 2.36 |

| | | |
|---|---|---|
| 1z62 | IAA | 1.9 |
| 3g2h | KOT | 2.03 |
| 1ww2 | NBG | 1.9 |

Table B.3.85 PDB structures used to construct the CRANkS grids for run 2 of 10 structures for target PYGM

PYGM, 10 structures, Run 3

| PDB Code | Ligand Residue Code | Resolution / Å |
|---|---|---|
| 3mrt | 12E | 1.98 |
| 3mrx | 17S | 1.95 |
| 1z8d | ADE | 2.3 |
| 2qln | F59 | 2.15 |
| 2g9r | G27 | 2.07 |
| 1xc7 | GL6 | 1.83 |
| 5gpb | GPM | 2.3 |
| 2qrm | M09 | 1.9 |
| 3g2i | RUG | 2 |
| 2qrp | S06 | 1.86 |

Table B.3.86 PDB structures used to construct the CRANkS grids for run 3 of 10 structures for target PYGM

PYGM, 10 structures, Run 4

| PDB Code | Ligand Residue Code | Resolution / Å |
|---|---|---|
| 3cuu | 376 | 2.3 |
| 3mrv | 16F | 1.94 |
| 2f3p | 4GP | 1.94 |
| 2f3q | 6GP | 1.96 |
| 3bd8 | C3B | 2.1 |
| 3l7b | DKZ | 2 |
| 2g9r | G27 | 2.07 |
| 4gpb | GFP | 2.3 |
| 2qrm | M09 | 1.9 |
| 3bcu | THM | 2.03 |

Table B.3.87 PDB structures used to construct the CRANkS grids for run 4 of 10 structures for target PYGM

PYGM, 10 structures, Run 5

| PDB Code | Ligand Residue Code | Resolution / Å |
|---|---|---|
| 3mrv | 16F | 1.94 |
| 3mt8 | 17T | 2 |
| 3mqf | 20X | 1.951 |
| 3nc4 | 26O | 2.07 |
| 2f3p | 4GP | 1.94 |
| 1p4j | CBF | 2 |
| 1c8l | CFF | 2.3 |
| 1z62 | IAA | 1.9 |
| 3g2l | LEW | 2.3 |

| | | |
|---|---|---|
| 3g2n | OAK | 2.1 |

Table B.3.88 PDB structures used to construct the CRANkS grids for run 5 of 10 structures for target PYGM

PYGM, 25 structures, Run 1

| PDB Code | Ligand Residue Code | Resolution / Å |
|---|---|---|
| 3mt8 | 17T | 2 |
| 3ebo | 57D | 1.9 |
| 2f3q | 6GP | 1.96 |
| 2f3u | 8GP | 1.93 |
| 1z8d | ADE | 2.3 |
| 1c8l | CFF | 2.3 |
| 1c50 | CHI | 2.3 |
| 3bd7 | CKB | 1.9 |
| 1fu8 | CR6 | 2.35 |
| 3l7c | DK4 | 1.93 |
| 3l7d | DK5 | 2 |
| 2qln | F59 | 2.15 |
| 2g9r | G27 | 2.07 |
| 4gpb | GFP | 2.3 |
| 1fty | GL7 | 2.38 |
| 1a8i | GLS | 1.78 |
| 2fet | H53 | 2.03 |
| 3g2h | KOT | 2.03 |
| 2qrg | M07 | 1.85 |
| 1ww2 | NBG | 1.9 |
| 2prj | NBG | 2.3 |
| 1ww3 | NTF | 1.8 |
| 1noi | NTZ | 2.5 |
| 2qrq | S13 | 1.8 |
| 3g2k | SKY | 2 |

Table B.3.89 PDB structures used to construct the CRANkS grids for run 1 of 25 structures for target PYGM

PYGM, 25 structures, Run 2

| PDB Code | Ligand Residue Code | Resolution / Å |
|---|---|---|
| 3cut | 179 | 2.3 |
| 3cuw | 445 | 2 |
| 3mrx | 17S | 1.95 |
| 2g9q | 1AB | 2.5 |
| 3mtd | 25E | 2.096 |
| 3nc4 | 26O | 2.07 |
| 2f3p | 4GP | 1.94 |
| 2f3q | 6GP | 1.96 |
| 3g2j | 9GP | 2.14 |
| 3bda | C4B | 2 |
| 3bcs | CJB | 2 |
| 3bd7 | CKB | 1.9 |

| | | |
|---|---|---|
| 3ebp | CPB | 2 |
| 1fu7 | CR1 | 2.36 |
| 3l7c | DK4 | 1.93 |
| 3l7a | DKY | 1.9 |
| 1ftw | GL5 | 2.36 |
| 1gg8 | GLG | 2.31 |
| 1z62 | IAA | 1.9 |
| 3g2h | KOT | 2.03 |
| 2qrm | M09 | 1.9 |
| 1ww2 | NBG | 1.9 |
| 1ww3 | NTF | 1.8 |
| 1xl0 | OX2 | 1.92 |
| 3g2i | RUG | 2 |

Table B.3.90 PDB structures used to construct the CRANkS grids for run 2 of 25 structures for target PYGM

PYGM, 25 structures, Run 3

| PDB Code | Ligand Residue Code | Resolution / Å |
|---|---|---|
| 3cuw | 445 | 2 |
| 3mrt | 12E | 1.98 |
| 3mrx | 17S | 1.95 |
| 3ms4 | 21N | 2.07 |
| 2f3s | 7GP | 1.96 |
| 1z8d | ADE | 2.3 |
| 6gpb | AMP | 2.86 |
| 3bda | C4B | 2 |
| 1p4j | CBF | 2 |
| 3bcs | CJB | 2 |
| 3bd7 | CKB | 1.9 |
| 1fu7 | CR1 | 2.36 |
| 3l79 | DKX | 1.86 |
| 3l7b | DKZ | 2 |
| 2ffr | DL6 | 2.03 |
| 2qln | F59 | 2.15 |
| 2g9r | G27 | 2.07 |
| 1xc7 | GL6 | 1.83 |
| 1a8i | GLS | 1.78 |
| 5gpb | GPM | 2.3 |
| 2qrm | M09 | 1.9 |
| 1ww3 | NTF | 1.8 |
| 3g2i | RUG | 2 |
| 2qrp | S06 | 1.86 |
| 3bcu | THM | 2.03 |

Table B.3.91 PDB structures used to construct the CRANkS grids for run 3 of 25 structures for target PYGM

PYGM, 25 structures, Run 4

| PDB Code | Ligand Residue Code | Resolution / Å |
|---|---|---|
| 3cuu | 376 | 2.3 |
| 3cuw | 445 | 2 |
| 3mrv | 16F | 1.94 |
| 3mqf | 20X | 1.951 |
| 2f3p | 4GP | 1.94 |
| 2f3q | 6GP | 1.96 |
| 1z8d | ADE | 2.3 |
| 3bd8 | C3B | 2.1 |
| 1fu8 | CR6 | 2.35 |
| 3l79 | DKX | 1.86 |
| 3l7b | DKZ | 2 |
| 3gpb | G1P | 2.3 |
| 2g9r | G27 | 2.07 |
| 4gpb | GFP | 2.3 |
| 1ftq | GL2 | 2.35 |
| 1hlf | GL4 | 2.26 |
| 1xc7 | GL6 | 1.83 |
| 1gg8 | GLG | 2.31 |
| 5gpb | GPM | 2.3 |
| 2fet | H53 | 2.03 |
| 2qrm | M09 | 1.9 |
| 2prj | NBG | 2.3 |
| 2qn9 | NBX | 2 |
| 3g2n | OAK | 2.1 |
| 3bcu | THM | 2.03 |

Table B.3.92 PDB structures used to construct the CRANkS grids for run 4 of 25 structures for target PYGM

PYGM, 25 structures, Run 5

| PDB Code | Ligand Residue Code | Resolution / Å |
|---|---|---|
| 3mrv | 16F | 1.94 |
| 3mrx | 17S | 1.95 |
| 3mt8 | 17T | 2 |
| 3mqf | 20X | 1.951 |
| 3ms7 | 22S | 1.95 |
| 3msc | 24S | 1.951 |
| 3mtd | 25E | 2.096 |
| 3nc4 | 26O | 2.07 |
| 2f3p | 4GP | 1.94 |
| 2f3s | 7GP | 1.96 |
| 1z8d | ADE | 2.3 |
| 1p4j | CBF | 2 |

| | | |
|---|---|---|
| 1c8l | CFF | 2.3 |
| 3l7c | DK4 | 1.93 |
| 3l7a | DKY | 1.9 |
| 2g9r | G27 | 2.07 |
| 1ftq | GL2 | 2.35 |
| 1xc7 | GL6 | 1.83 |
| 1a8i | GLS | 1.78 |
| 1z62 | IAA | 1.9 |
| 3g2l | LEW | 2.3 |
| 2qrh | M08 | 1.83 |
| 1ww2 | NBG | 1.9 |
| 2qn8 | NBY | 1.9 |
| 3g2n | OAK | 2.1 |

Table B.3.93 PDB structures used to construct the CRANkS grids for run 5 of 25 structures for target PYGM

PYGM, 50 structures, Run 1

| PDB Code | Ligand Residue Code | Resolution / Å |
|---|---|---|
| 3cut | 179 | 2.3 |
| 3cuu | 376 | 2.3 |
| 3mt7 | 16O | 2 |
| 3mrx | 17S | 1.95 |
| 3mt8 | 17T | 2 |
| 3mt9 | 18O | 2.05 |
| 3mtb | 23V | 1.95 |
| 2f3p | 4GP | 1.94 |
| 3ebo | 57D | 1.9 |
| 2f3q | 6GP | 1.96 |
| 2f3u | 8GP | 1.93 |
| 3g2j | 9GP | 2.14 |
| 1z8d | ADE | 2.3 |
| 6gpb | AMP | 2.86 |
| 3bda | C4B | 2 |
| 1c8l | CFF | 2.3 |
| 1c50 | CHI | 2.3 |
| 3bd7 | CKB | 1.9 |
| 3ebp | CPB | 2 |
| 1fu8 | CR6 | 2.35 |
| 1b4d | CRA | 2 |
| 3l7c | DK4 | 1.93 |
| 3l7d | DK5 | 2 |
| 3l7a | DKY | 1.9 |
| 3l7b | DKZ | 2 |
| 2qn3 | F55 | 1.96 |
| 2qln | F59 | 2.15 |
| 3gpb | G1P | 2.3 |
| 2g9r | G27 | 2.07 |
| 4gpb | GFP | 2.3 |
| 1ftw | GL5 | 2.36 |

| | | |
|---|---|---|
| 1xc7 | GL6 | 1.83 |
| 1fty | GL7 | 2.38 |
| 1h5u | GLC | 1.76 |
| 1gg8 | GLG | 2.31 |
| 1a8i | GLS | 1.78 |
| 2fet | H53 | 2.03 |
| 3g2h | KOT | 2.03 |
| 3g2l | LEW | 2.3 |
| 2qrg | M07 | 1.85 |
| 2qrh | M08 | 1.83 |
| 1ww2 | NBG | 1.9 |
| 2prj | NBG | 2.3 |
| 2qn9 | NBX | 2 |
| 1ww3 | NTF | 1.8 |
| 1noi | NTZ | 2.5 |
| 1xl0 | OX2 | 1.92 |
| 2qrp | S06 | 1.86 |
| 2qrq | S13 | 1.8 |
| 3g2k | SKY | 2 |

Table B.3.94 PDB structures used to construct the CRANkS grids for run 1 of 50 structures for target PYGM

PYGM, 50 structures, Run 2

| PDB Code | Ligand Residue Code | Resolution / Å |
|---|---|---|
| 3cut | 179 | 2.3 |
| 3cuu | 376 | 2.3 |
| 3cuw | 445 | 2 |
| 3cuv | 475 | 1.93 |
| 3mrx | 17S | 1.95 |
| 3ms2 | 18S | 2.1 |
| 2g9q | 1AB | 2.5 |
| 3ms7 | 22S | 1.95 |
| 3mtb | 23V | 1.95 |
| 3mtd | 25E | 2.096 |
| 3nc4 | 26O | 2.07 |
| 2f3p | 4GP | 1.94 |
| 3ebo | 57D | 1.9 |
| 2f3q | 6GP | 1.96 |
| 3g2j | 9GP | 2.14 |
| 1z8d | ADE | 2.3 |
| 1kti | AZC | 1.97 |
| 3bda | C4B | 2 |
| 1p4j | CBF | 2 |
| 1c50 | CHI | 2.3 |
| 3bcs | CJB | 2 |
| 3bd7 | CKB | 1.9 |
| 3ebp | CPB | 2 |
| 1fu7 | CR1 | 2.36 |
| 1b4d | CRA | 2 |
| 3l7c | DK4 | 1.93 |
| 3l7d | DK5 | 2 |

| | | |
|---|---|---|
| 3l79 | DKX | 1.86 |
| 3l7a | DKY | 1.9 |
| 3l7b | DKZ | 2 |
| 2ffr | DL6 | 2.03 |
| 2qln | F59 | 2.15 |
| 1hlf | GL4 | 2.26 |
| 1ftw | GL5 | 2.36 |
| 1xc7 | GL6 | 1.83 |
| 1h5u | GLC | 1.76 |
| 1gg8 | GLG | 2.31 |
| 1z62 | IAA | 1.9 |
| 2g9v | IFM | 2.15 |
| 3g2h | KOT | 2.03 |
| 3g2l | LEW | 2.3 |
| 2qrh | M08 | 1.83 |
| 2qrm | M09 | 1.9 |
| 1ww2 | NBG | 1.9 |
| 2prj | NBG | 2.3 |
| 2qn8 | NBY | 1.9 |
| 1ww3 | NTF | 1.8 |
| 1xlo | OX2 | 1.92 |
| 3g2i | RUG | 2 |
| 3bcu | THM | 2.03 |

Table B.3.95 PDB structures used to construct the CRANkS grids for run 2 of 50 structures for target PYGM

PYGM, 50 structures, Run 3

| PDB Code | Ligand Residue Code | Resolution / Å |
|---|---|---|
| 3cuu | 376 | 2.3 |
| 3cuw | 445 | 2 |
| 3mrt | 12E | 1.98 |
| 3mrx | 17S | 1.95 |
| 2g9q | 1AB | 2.5 |
| 3ms4 | 21N | 2.07 |
| 3mtb | 23V | 1.95 |
| 3mtd | 25E | 2.096 |
| 3nc4 | 26O | 2.07 |
| 2f3p | 4GP | 1.94 |
| 2f3s | 7GP | 1.96 |
| 1z8d | ADE | 2.3 |
| 6gpb | AMP | 2.86 |
| 2qnb | BZD | 1.8 |
| 3bda | C4B | 2 |
| 1p4j | CBF | 2 |
| 3bcs | CJB | 2 |
| 3bd7 | CKB | 1.9 |
| 1fu7 | CR1 | 2.36 |
| 1b4d | CRA | 2 |
| 3l7c | DK4 | 1.93 |
| 3l7d | DK5 | 2 |
| 3l79 | DKX | 1.86 |

| | | |
|---|---|---|
| 3l7a | DKY | 1.9 |
| 3l7b | DKZ | 2 |
| 2ffr | DL6 | 2.03 |
| 2qn3 | F55 | 1.96 |
| 2qln | F59 | 2.15 |
| 2g9r | G27 | 2.07 |
| 4gpb | GFP | 2.3 |
| 1ftw | GL5 | 2.36 |
| 1xc7 | GL6 | 1.83 |
| 1h5u | GLC | 1.76 |
| 1gg8 | GLG | 2.31 |
| 1a8i | GLS | 1.78 |
| 5gpb | GPM | 2.3 |
| 2fet | H53 | 2.03 |
| 1z62 | IAA | 1.9 |
| 3g2h | KOT | 2.03 |
| 3g2l | LEW | 2.3 |
| 2qrm | M09 | 1.9 |
| 2qn8 | NBY | 1.9 |
| 1ww3 | NTF | 1.8 |
| 1noi | NTZ | 2.5 |
| 3g2n | OAK | 2.1 |
| 1xlo | OX2 | 1.92 |
| 3g2i | RUG | 2 |
| 2qrp | S06 | 1.86 |
| 3g2k | SKY | 2 |
| 3bcu | THM | 2.03 |

Table B.3.96 PDB structures used to construct the CRANkS grids for run 3 of 50 structures for target PYGM

PYGM, 50 structures, Run 4

| PDB Code | Ligand Residue Code | Resolution / Å |
|---|---|---|
| 3cuu | 376 | 2.3 |
| 3cuw | 445 | 2 |
| 3cuv | 475 | 1.93 |
| 3mrv | 16F | 1.94 |
| 3mrx | 17S | 1.95 |
| 3mt8 | 17T | 2 |
| 2g9q | 1AB | 2.5 |
| 3mqf | 20X | 1.951 |
| 3ms4 | 21N | 2.07 |
| 3mtd | 25E | 2.096 |
| 3nc4 | 26O | 2.07 |
| 2f3p | 4GP | 1.94 |
| 2f3q | 6GP | 1.96 |
| 2f3u | 8GP | 1.93 |
| 3g2j | 9GP | 2.14 |
| 1z8d | ADE | 2.3 |
| 1kti | AZC | 1.97 |
| 2qnb | BZD | 1.8 |

| | | |
|---|---|---|
| 3bd8 | C3B | 2.1 |
| 1p4j | CBF | 2 |
| 1c50 | CHI | 2.3 |
| 3bd7 | CKB | 1.9 |
| 1fu8 | CR6 | 2.35 |
| 3l7d | DK5 | 2 |
| 3l79 | DKX | 1.86 |
| 3l7b | DKZ | 2 |
| 2qn3 | F55 | 1.96 |
| 2qln | F59 | 2.15 |
| 3gpb | G1P | 2.3 |
| 2g9r | G27 | 2.07 |
| 4gpb | GFP | 2.3 |
| 1ftq | GL2 | 2.35 |
| 1hlf | GL4 | 2.26 |
| 1xc7 | GL6 | 1.83 |
| 1fty | GL7 | 2.38 |
| 1gg8 | GLG | 2.31 |
| 5gpb | GPM | 2.3 |
| 2fet | H53 | 2.03 |
| 1z62 | IAA | 1.9 |
| 3g2l | LEW | 2.3 |
| 2qrm | M09 | 1.9 |
| 1ww2 | NBG | 1.9 |
| 2prj | NBG | 2.3 |
| 2qn9 | NBX | 2 |
| 2qn8 | NBY | 1.9 |
| 1noi | NTZ | 2.5 |
| 3g2n | OAK | 2.1 |
| 1xlo | OX2 | 1.92 |
| 2qrq | S13 | 1.8 |
| 3bcu | THM | 2.03 |

Table B.3.97 PDB structures used to construct the CRANkS grids for run 4 of 50 structures for target PYGM

PYGM, 50 structures, Run 5

| PDB Code | Ligand Residue Code | Resolution / Å |
|---|---|---|
| 3cuv | 475 | 1.93 |
| 3mrv | 16F | 1.94 |
| 3mrx | 17S | 1.95 |
| 3mt8 | 17T | 2 |
| 3mqf | 20X | 1.951 |
| 3ms4 | 21N | 2.07 |
| 3ms7 | 22S | 1.95 |
| 3msc | 24S | 1.951 |
| 3mtd | 25E | 2.096 |
| 3nc4 | 26O | 2.07 |
| 2f3p | 4GP | 1.94 |
| 2f3s | 7GP | 1.96 |
| 2f3u | 8GP | 1.93 |

| | | |
|---|---|---|
| 3g2j | 9GP | 2.14 |
| 1z8d | ADE | 2.3 |
| 6gpb | AMP | 2.86 |
| 1kti | AZC | 1.97 |
| 3bcr | AZZ | 2.14 |
| 2qnb | BZD | 1.8 |
| 1p4j | CBF | 2 |
| 1c8l | CFF | 2.3 |
| 1c50 | CHI | 2.3 |
| 3bcs | CJB | 2 |
| 1fu7 | CR1 | 2.36 |
| 1fu8 | CR6 | 2.35 |
| 3l7c | DK4 | 1.93 |
| 3l7d | DK5 | 2 |
| 3l7a | DKY | 1.9 |
| 3l7b | DKZ | 2 |
| 2ffr | DL6 | 2.03 |
| 2qn3 | F55 | 1.96 |
| 3gpb | G1P | 2.3 |
| 2g9r | G27 | 2.07 |
| 4gpb | GFP | 2.3 |
| 1ftq | GL2 | 2.35 |
| 1xc7 | GL6 | 1.83 |
| 1fu4 | GL9 | 2.36 |
| 1a8i | GLS | 1.78 |
| 5gpb | GPM | 2.3 |
| 1z62 | IAA | 1.9 |
| 3g2l | LEW | 2.3 |
| 2qrg | M07 | 1.85 |
| 2qrh | M08 | 1.83 |
| 1ww2 | NBG | 1.9 |
| 2prj | NBG | 2.3 |
| 2qn8 | NBY | 1.9 |
| 1noi | NTZ | 2.5 |
| 3g2n | OAK | 2.1 |
| 1xlo | OX2 | 1.92 |
| 3bcu | THM | 2.03 |

Table B.3.98 PDB structures used to construct the CRANkS grids for run 5 of 50 structures for target PYGM

CDK2, 5 structures, Run 1

| PDB Code | Ligand Residue Code | Resolution / Å |
|---|---|---|
| 1pxi | CK1 | 1.95 |
| 1di8 | DTQ | 2.2 |
| 1oit | HDT | 1.6 |
| 2c5y | MTW | 2.25 |
| 1pye | PM1 | 2 |

Table B.3.99 PDB structures used to construct the CRANkS grids for run 1 of 5 structures for target CDK2

CDK2, 5 structures, Run 2

| PDB Code | Ligand Residue Code | Resolution / Å |
|---|---|---|
| 3ej1 | 5BP | 3.22 |
| 1pxi | CK1 | 1.95 |
| 1r78 | FMD | 2 |
| 2uue | MTZ | 2.06 |
| 1pye | PM1 | 2 |

Table B.3.100 PDB structures used to construct the CRANkS grids for run 2 of 5 structures for target CDK2

CDK2, 5 structures, Run 3

| PDB Code | Ligand Residue Code | Resolution / Å |
|---|---|---|
| 2r64 | 740 | 2.3 |
| 3pxf | 2AN | 1.8 |
| 2wih | P48 | 2.5 |
| 1jsv | U55 | 1.96 |
| 2btr | U73 | 1.85 |

Table B.3.101 PDB structures used to construct the CRANkS grids for run 3 of 5 structures for target CDK2

CDK2, 5 structures, Run 4

| PDB Code | Ligand Residue Code | Resolution / Å |
|---|---|---|
| 2uze | C95 | 2.4 |
| 2b53 | D23 | 2 |
| 2w17 | I19 | 2.15 |
| 1ke6 | LS2 | 2 |
| 1ke9 | LS5 | 2 |

Table B.3.102 PDB structures used to construct the CRANkS grids for run 4 of 5 structures for target CDK2

CDK2, 5 structures, Run 5

| PDB Code | Ligand Residue Code | Resolution / Å |
|---|---|---|
| 1gih | 1PU | 2.8 |
| 1gij | 2PU | 2.2 |
| 3lfq | A28 | 2.03 |
| 2uze | C95 | 2.4 |
| 2c6k | DT2 | 1.9 |

Table B.3.103 PDB structures used to construct the CRANkS grids for run 5 of 5 structures for target CDK2

CDK2, 10 structures, Run 1

| PDB Code | Ligand Residue Code | Resolution / Å |
|---|---|---|
| 1pxi | CK1 | 1.95 |

| | | |
|---|---|---|
| 2wev | CK7 | 2.3 |
| 2c6l | DT4 | 2.3 |
| 1di8 | DTQ | 2.2 |
| 1h00 | FAP | 1.6 |
| 1h00 | FCP | 1.6 |
| 1oit | HDT | 1.6 |
| 2w1h | LoF | 2.15 |
| 2c5y | MTW | 2.25 |
| 1vyz | N5B | 2.21 |
| 1pye | PM1 | 2 |

Table B.3.104 PDB structures used to construct the CRANkS grids for run 1 of 10 structures for target CDK2

CDK2, 10 structures, Run 2

| PDB Code | Ligand Residue Code | Resolution / Å |
|---|---|---|
| 2duv | 371 | 2.2 |
| 3le6 | 2BZ | 2 |
| 3ej1 | 5BP | 3.22 |
| 1pxi | CK1 | 1.95 |
| 3ezv | EZV | 1.99 |
| 1r78 | FMD | 2 |
| 2vtj | LZ4 | 2.2 |
| 2uue | MTZ | 2.06 |
| 1w8c | N69 | 2.05 |
| 1pye | PM1 | 2 |

Table B.3.105 PDB structures used to construct the CRANkS grids for run 2 of 10 structures for target CDK2

CDK2, 10 structures, Run 3

| PDB Code | Ligand Residue Code | Resolution / Å |
|---|---|---|
| 2r64 | 740 | 2.3 |
| 3pxf | 2AN | 1.8 |
| 1pxp | CK8 | 2.3 |
| 3pxz | JWS | 1.7 |
| 1ke6 | LS2 | 2 |
| 2vto | LZ8 | 2.19 |
| 3bht | MFR | 2 |
| 2wih | P48 | 2.5 |
| 1jsv | U55 | 1.96 |
| 2btr | U73 | 1.85 |

Table B.3.106 PDB structures used to construct the CRANkS grids for run 3 of 10 structures for target CDK2

CDK2, 10 structures, Run 4

| PDB Code | Ligand Residue Code | Resolution / Å |
|---|---|---|
| 1how | 207 | 2.1 |
| 1gih | 1PU | 2.8 |
| 2uzl | C94 | 2.4 |
| 2uze | C95 | 2.4 |
| 1pxi | CK1 | 1.95 |
| 2b53 | D23 | 2 |
| 2w17 | I19 | 2.15 |
| 1ke6 | LS2 | 2 |
| 1ke9 | LS5 | 2 |
| 2wip | P49 | 2.8 |

Table B.3.107 PDB structures used to construct the CRANkS grids for run 4 of 10 structures for target CDK2

CDK2, 10 structures, Run 5

| PDB Code | Ligand Residue Code | Resolution / Å |
|---|---|---|
| 1gih | 1PU | 2.8 |
| 1gij | 2PU | 2.2 |
| 3lfq | A28 | 2.03 |
| 2uzb | C75 | 2.7 |
| 2uze | C95 | 2.4 |
| 2c6k | DT2 | 1.9 |
| 2c6l | DT4 | 2.3 |
| 1oi9 | N20 | 2.1 |
| 3dog | NNN | 2.7 |
| 1aq1 | STU | 2 |

Table B.3.108 PDB structures used to construct the CRANkS grids for run 5 of 10 structures for target CDK2

CDK2, 25 structures, Run 1

| PDB Code | Ligand Residue Code | Resolution / Å |
|---|---|---|
| 1gij | 2PU | 2.2 |
| 1wcc | CIG | 2.2 |
| 1pxi | CK1 | 1.95 |
| 2c5v | CK4 | 2.9 |
| 2wev | CK7 | 2.3 |
| 1pxp | CK8 | 2.3 |
| 2c6l | DT4 | 2.3 |
| 1di8 | DTQ | 2.2 |
| 3igg | EFQ | 1.8 |
| 1hoo | FAP | 1.6 |
| 1hoo | FCP | 1.6 |

| | | |
|---|---|---|
| 2w06 | FRV | 2.04 |
| 1oit | HDT | 1.6 |
| 2w1h | LoF | 2.15 |
| 1ke7 | LS3 | 2 |
| 1ke8 | LS4 | 2 |
| 2vtq | LZA | 1.9 |
| 1h07 | MFP | 1.85 |
| 2c5y | MTW | 2.25 |
| 1vyz | N5B | 2.21 |
| 3ns9 | NS9 | 1.78 |
| 1pye | PM1 | 2 |
| 2bkz | SBC | 2.6 |
| 2r3h | SCE | 1.5 |
| 1ogu | ST8 | 2.6 |
| 2wxv | WXV | 2.6 |

Table B.3.109 PDB structures used to construct the CRANkS grids for run 1 of 25 structures for target CDK2

CDK2, 25 structures, Run 2

| PDB Code | Ligand Residue Code | Resolution / Å |
|---|---|---|
| 1fvt | 106 | 2.2 |
| 1vyw | 292 | 2.3 |
| 2duv | 371 | 2.2 |
| 3le6 | 2BZ | 2 |
| 3ej1 | 5BP | 3.22 |
| 2uzb | C75 | 2.7 |
| 2uzl | C94 | 2.4 |
| 1pxi | CK1 | 1.95 |
| 1pxm | CK5 | 2.53 |
| 2b52 | D42 | 1.88 |
| 1di8 | DTQ | 2.2 |
| 3igg | EFQ | 1.8 |
| 3ezv | EZV | 1.99 |
| 2clx | F18 | 1.8 |
| 1r78 | FMD | 2 |
| 1g5s | I17 | 2.61 |
| 2vtj | LZ4 | 2.2 |
| 1gz8 | MBP | 1.3 |
| 2uue | MTZ | 2.06 |
| 1w8c | N69 | 2.05 |
| 2wih | P48 | 2.5 |
| 1pye | PM1 | 2 |
| 2bkz | SBC | 2.6 |
| 1p5e | TBS | 2.22 |
| 2bts | U32 | 1.99 |

Table B.3.110 PDB structures used to construct the CRANkS grids for run 2 of 25 structures for target CDK2

CDK2, 25 structures, Run 3

| PDB Code | Ligand Residue Code | Resolution / Å |
|---|---|---|
| 2duv | 37I | 2.2 |
| 2c4g | 5I4 | 2.7 |
| 2r64 | 740 | 2.3 |
| 3pxf | 2AN | 1.8 |
| 1b38 | ATP | 2 |
| 1pxk | CK3 | 2.8 |
| 1pxp | CK8 | 2.3 |
| 2a0c | CK9 | 1.95 |
| 2b52 | D42 | 1.88 |
| 2c6k | DT2 | 1.9 |
| 3ig7 | EFP | 1.8 |
| 3ezv | EZV | 1.99 |
| 1h01 | FAL | 1.79 |
| 1h01 | FBL | 1.79 |
| 1oir | HDY | 1.91 |
| 2w17 | I19 | 2.15 |
| 3pxz | JWS | 1.7 |
| 1ke6 | LS2 | 2 |
| 2vto | LZ8 | 2.19 |
| 1gz8 | MBP | 1.3 |
| 3bht | MFR | 2 |
| 2wih | P48 | 2.5 |
| 2bhh | RYU | 2.6 |
| 1jsv | U55 | 1.96 |
| 2btr | U73 | 1.85 |
| 2exm | ZIP | 1.8 |

Table B.3.111 PDB structures used to construct the CRANkS grids for run 3 of 25 structures for target CDK2

CDK2, 25 structures, Run 4

| PDB Code | Ligand Residue Code | Resolution / Å |
|---|---|---|
| 1how | 207 | 2.1 |
| 2duv | 37I | 2.2 |
| 2wpa | 889 | 2.5I |
| 1gih | 1PU | 2.8 |
| 3lfq | A28 | 2.03 |
| 2uzl | C94 | 2.4 |
| 2uze | C95 | 2.4 |
| 1pxi | CK1 | 1.95 |
| 2wev | CK7 | 2.3 |
| 2b53 | D23 | 2 |
| 1h01 | FAL | 1.79 |
| 1hoo | FAP | 1.6 |
| 1h01 | FBL | 1.79 |
| 1hoo | FCP | 1.6 |

| | | |
|---|---|---|
| 1oir | HDY | 1.91 |
| 2wi7 | I19 | 2.15 |
| 2fvd | LIA | 1.85 |
| 1ke6 | LS2 | 2 |
| 1ke9 | LS5 | 2 |
| 2vth | LZ2 | 1.9 |
| 2vto | LZ8 | 2.19 |
| 3bhu | MHR | 2.3 |
| 2uue | MTZ | 2.06 |
| 2wih | P48 | 2.5 |
| 2wip | P49 | 2.8 |
| 3ddp | RC8 | 2.7 |
| 2exm | ZIP | 1.8 |

Table B.3.112 PDB structures used to construct the CRANkS grids for run 4 of 25 structures for target CDK2

CDK2, 25 structures, Run 5

| PDB Code | Ligand Residue Code | Resolution / Å |
|---|---|---|
| 1gih | 1PU | 2.8 |
| 3pxq | 2AN | 1.9 |
| 1gij | 2PU | 2.2 |
| 1v1k | 3FP | 2.31 |
| 3lfq | A28 | 2.03 |
| 3fz1 | B98 | 1.9 |
| 1h08 | BWP | 1.8 |
| 1h08 | BYP | 1.8 |
| 2uzo | C62 | 2.3 |
| 2uzb | C75 | 2.7 |
| 2uze | C95 | 2.4 |
| 1pxl | CK4 | 2.5 |
| 1pxp | CK8 | 2.3 |
| 2c6k | DT2 | 1.9 |
| 2c6l | DT4 | 2.3 |
| 1g5s | I17 | 2.61 |
| 1ke9 | LS5 | 2 |
| 1h07 | MFP | 1.85 |
| 1oi9 | N20 | 2.1 |
| 3dog | NNN | 2.7 |
| 3ns9 | NS9 | 1.78 |
| 1wox | OLO | 2.2 |
| 3eid | PO5 | 3.15 |
| 2a4l | RRC | 2.4 |
| 1aq1 | STU | 2 |
| 1pf8 | SU9 | 2.51 |

Table B.3.113 PDB structures used to construct the CRANkS grids for run 5 of 25 structures for target CDK2

CDK2, 50 structures, Run 1

| PDB Code | Ligand Residue Code | Resolution / Å |
|---|---|---|
| 1fvt | 106 | 2.2 |
| 1how | 207 | 2.1 |
| 3pxf | 2AN | 1.8 |
| 1gij | 2PU | 2.2 |
| 1v1k | 3FP | 2.31 |
| 1ho8 | BWP | 1.8 |
| 1ho8 | BYP | 1.8 |
| 2uze | C95 | 2.4 |
| 2uzn | C96 | 2.3 |
| 1wcc | CIG | 2.2 |
| 1pxi | CK1 | 1.95 |
| 2c5v | CK4 | 2.9 |
| 2wev | CK7 | 2.3 |
| 1pxp | CK8 | 2.3 |
| 2a0c | CK9 | 1.95 |
| 1e1v | CMG | 1.95 |
| 2c6l | DT4 | 2.3 |
| 1di8 | DTQ | 2.2 |
| 3igg | EFQ | 1.8 |
| 3ezv | EZV | 1.99 |
| 1hoo | FAP | 1.6 |
| 1hoo | FCP | 1.6 |
| 1r78 | FMD | 2 |
| 2w06 | FRV | 2.04 |
| 1oit | HDT | 1.6 |
| 2w17 | I19 | 2.15 |
| 1urw | I1P | 1.6 |
| 1e9h | INR | 2.5 |
| 2w1h | L0F | 2.15 |
| 1ke7 | LS3 | 2 |
| 1ke8 | LS4 | 2 |
| 2vto | LZ8 | 2.19 |
| 2vtq | LZA | 1.9 |
| 2vu3 | LZE | 1.85 |
| 1gz8 | MBP | 1.3 |
| 1ho7 | MFP | 1.85 |
| 2c5y | MTW | 2.25 |
| 1vyz | N5B | 2.21 |
| 3dog | NNN | 2.7 |
| 3ns9 | NS9 | 1.78 |
| 2g9x | NU5 | 2.5 |
| 1wox | OLO | 2.2 |
| 1pye | PM1 | 2 |
| 3ddp | RC8 | 2.7 |
| 3my5 | RFZ | 2.1 |
| 2bkz | SBC | 2.6 |
| 2r3h | SCE | 1.5 |
| 1ogu | ST8 | 2.6 |

| | | |
|---|---|---|
| 3eoc | T2A | 3.2 |
| 1pkd | UCN | 2.3 |
| 1hov | UN4 | 1.9 |
| 2wxv | WXV | 2.6 |

Table B.3.114 PDB structures used to construct the CRANkS grids for run 1 of 50 structures for target CDK2

CDK2, 50 structures, Run 2

| PDB Code | Ligand Residue Code | Resolution / Å |
|---|---|---|
| 1fvt | 106 | 2.2 |
| 1vyw | 292 | 2.3 |
| 2duv | 371 | 2.2 |
| 2c4g | 514 | 2.7 |
| 3le6 | 2BZ | 2 |
| 2c60 | 4SP | 2.1 |
| 3ej1 | 5BP | 3.22 |
| 3fz1 | B98 | 1.9 |
| 2bhe | BRY | 1.9 |
| 2uzb | C75 | 2.7 |
| 2uzl | C94 | 2.4 |
| 2uzn | C96 | 2.3 |
| 1pxi | CK1 | 1.95 |
| 1pxm | CK5 | 2.53 |
| 2wev | CK7 | 2.3 |
| 1h1p | CMG | 2.1 |
| 2b55 | D31 | 1.85 |
| 2b52 | D42 | 1.88 |
| 2c6t | DT5 | 2.61 |
| 1di8 | DTQ | 2.2 |
| 3igg | EFQ | 1.8 |
| 3ezv | EZV | 1.99 |
| 2clx | F18 | 1.8 |
| 1r78 | FMD | 2 |
| 2w06 | FRV | 2.04 |
| 1g5s | I17 | 2.61 |
| 2w17 | I19 | 2.15 |
| 2vv9 | IM9 | 1.9 |
| 2w1h | LoF | 2.15 |
| 2fvd | LIA | 1.85 |
| 2vti | LZ3 | 2 |
| 2vtj | LZ4 | 2.2 |
| 2vto | LZ8 | 2.19 |
| 2vtp | LZ9 | 2.15 |
| 2vtt | LZD | 1.68 |
| 2vu3 | LZE | 1.85 |
| 1gz8 | MBP | 1.3 |
| 2uue | MTZ | 2.06 |
| 1w8c | N69 | 2.05 |
| 3dog | NNN | 2.7 |
| 1wox | OLO | 2.2 |

| | | |
|---|---|---|
| 2wih | P48 | 2.5 |
| 2wip | P49 | 2.8 |
| 1pye | PM1 | 2 |
| 2iw6 | QQ2 | 2.3 |
| 3ddq | RRC | 1.8 |
| 2bkz | SBC | 2.6 |
| 1p5e | TBS | 2.22 |
| 2bts | U32 | 1.99 |
| 2x1n | X1N | 2.75 |

Table B.3.115 PDB structures used to construct the CRANkS grids for run 2 of 50 structures for target CDK2

CDK2, 50 structures, Run 3

| PDB Code | Ligand Residue Code | Resolution / Å |
|---|---|---|
| 1vyw | 292 | 2.3 |
| 2duv | 371 | 2.2 |
| 2c4g | 514 | 2.7 |
| 2r64 | 740 | 2.3 |
| 3pxf | 2AN | 1.8 |
| 3ej1 | 5BP | 3.22 |
| 1b38 | ATP | 2 |
| 2i40 | BLZ | 2.8 |
| 2uze | C95 | 2.4 |
| 1pxk | CK3 | 2.8 |
| 1pxp | CK8 | 2.3 |
| 2a0c | CK9 | 1.95 |
| 1e1v | CMG | 1.95 |
| 1y91 | CT9 | 2.15 |
| 2b52 | D42 | 1.88 |
| 2c6k | DT2 | 1.9 |
| 3ig7 | EFP | 1.8 |
| 3ezv | EZV | 1.99 |
| 2clx | F18 | 1.8 |
| 1h01 | FAL | 1.79 |
| 1h00 | FAP | 1.6 |
| 1h01 | FBL | 1.79 |
| 1h00 | FCP | 1.6 |
| 2w05 | FRT | 1.9 |
| 1oit | HDT | 1.6 |
| 1oir | HDY | 1.91 |
| 1dm2 | HMD | 2.1 |
| 2w17 | I19 | 2.15 |
| 2vv9 | IM9 | 1.9 |
| 1e9h | INR | 2.5 |
| 3pxz | JWS | 1.7 |
| 1ke6 | LS2 | 2 |
| 2vto | LZ8 | 2.19 |
| 2vtp | LZ9 | 2.15 |
| 2vtm | LZM | 2.25 |
| 1gz8 | MBP | 1.3 |

| | | |
|---|---|---|
| 3bht | MFR | 2 |
| 2c5y | MTW | 2.25 |
| 2uue | MTZ | 2.06 |
| 1w8c | N69 | 2.05 |
| 1oiu | N76 | 2 |
| 2wih | P48 | 2.5 |
| 3ddq | RRC | 1.8 |
| 2bhh | RYU | 2.6 |
| 2bkz | SBC | 2.6 |
| 2r3h | SCE | 1.5 |
| 1pf8 | SU9 | 2.51 |
| 1jsv | U55 | 1.96 |
| 2btr | U73 | 1.85 |
| 1pkd | UCN | 2.3 |
| 2wxv | WXV | 2.6 |
| 2exm | ZIP | 1.8 |

Table B.3.115 PDB structures used to construct the CRANkS grids for run 3 of 50 structures for target CDK2

CDK2, 50 structures, Run 4

| PDB Code | Ligand Residue Code | Resolution / Å |
|---|---|---|
| 1how | 207 | 2.1 |
| 2duv | 371 | 2.2 |
| 2wpa | 889 | 2.51 |
| 1gih | 1PU | 2.8 |
| 3le6 | 2BZ | 2 |
| 3ej1 | 5BP | 3.22 |
| 3lfq | A28 | 2.03 |
| 1b38 | ATP | 2 |
| 2uzo | C62 | 2.3 |
| 2uzb | C75 | 2.7 |
| 2uzl | C94 | 2.4 |
| 2uze | C95 | 2.4 |
| 1pxi | CK1 | 1.95 |
| 2c5v | CK4 | 2.9 |
| 2wev | CK7 | 2.3 |
| 2c69 | CT8 | 2.1 |
| 2b53 | D23 | 2 |
| 2c6i | DT1 | 1.8 |
| 2c6k | DT2 | 1.9 |
| 3igg | EFQ | 1.8 |
| 3ezv | EZV | 1.99 |
| 1ho1 | FAL | 1.79 |
| 1hoo | FAP | 1.6 |
| 1ho1 | FBL | 1.79 |
| 1hoo | FCP | 1.6 |
| 2w06 | FRV | 2.04 |
| 1oir | HDY | 1.91 |
| 2w17 | I19 | 2.15 |
| 3pxy | JWS | 1.8 |

| | | |
|---|---|---|
| 2w1h | LoF | 2.15 |
| 2fvd | LIA | 1.85 |
| 1ke6 | LS2 | 2 |
| 1ke9 | LS5 | 2 |
| 2vth | LZ2 | 1.9 |
| 2vto | LZ8 | 2.19 |
| 2vtq | LZA | 1.9 |
| 2vtm | LZM | 2.25 |
| 3bhu | MHR | 2.3 |
| 2uue | MTZ | 2.06 |
| 3dog | NNN | 2.7 |
| 2wih | P48 | 2.5 |
| 2wip | P49 | 2.8 |
| 2iw6 | QQ2 | 2.3 |
| 3ddp | RC8 | 2.7 |
| 2a4l | RRC | 2.4 |
| 2r3h | SCE | 1.5 |
| 1aq1 | STU | 2 |
| 3eoc | T2A | 3.2 |
| 1p5e | TBS | 2.22 |
| 1pkd | UCN | 2.3 |
| 2x1n | X1N | 2.75 |
| 2exm | ZIP | 1.8 |

Table B.3.116 PDB structures used to construct the CRANkS grids for run 4 of 50 structures for target CDK2

CDK2, 50 structures, Run 5

| PDB Code | Ligand Residue Code | Resolution / Å |
|---|---|---|
| 2bpm | 529 | 2.4 |
| 1ykr | 628 | 1.8 |
| 1gih | 1PU | 2.8 |
| 3pxq | 2AN | 1.9 |
| 1gij | 2PU | 2.2 |
| 1v1k | 3FP | 2.31 |
| 3lfq | A28 | 2.03 |
| 1b38 | ATP | 2 |
| 3fz1 | B98 | 1.9 |
| 2i40 | BLZ | 2.8 |
| 1h08 | BWP | 1.8 |
| 1h08 | BYP | 1.8 |
| 2v0d | C53 | 2.2 |
| 2uzo | C62 | 2.3 |
| 2uzb | C75 | 2.7 |
| 2uze | C95 | 2.4 |
| 1pxl | CK4 | 2.5 |
| 1pxn | CK6 | 2.5 |
| 2wev | CK7 | 2.3 |
| 1pxp | CK8 | 2.3 |
| 1y8y | CT7 | 1.996 |
| 2c6k | DT2 | 1.9 |

| | | |
|---|---|---|
| 2c6l | DT4 | 2.3 |
| 2c6t | DT5 | 2.61 |
| 3ig7 | EFP | 1.8 |
| 1h00 | FAP | 1.6 |
| 1h00 | FCP | 1.6 |
| 2w06 | FRV | 2.04 |
| 1dm2 | HMD | 2.1 |
| 1g5s | I17 | 2.61 |
| 2w17 | I19 | 2.15 |
| 3pxy | JWS | 1.8 |
| 1ke5 | LS1 | 2.2 |
| 1ke6 | LS2 | 2 |
| 1ke9 | LS5 | 2 |
| 2vta | LZ1 | 2 |
| 2vto | LZ8 | 2.19 |
| 1h07 | MFP | 1.85 |
| 1oi9 | N20 | 2.1 |
| 3dog | NNN | 2.7 |
| 3ns9 | NS9 | 1.78 |
| 2g9x | NU5 | 2.5 |
| 1w0x | OLO | 2.2 |
| 1pye | PM1 | 2 |
| 3eid | PO5 | 3.15 |
| 3ddp | RC8 | 2.7 |
| 2a4l | RRC | 2.4 |
| 1ogu | ST8 | 2.6 |
| 1aq1 | STU | 2 |
| 1pf8 | SU9 | 2.51 |
| 3eoc | T2A | 3.2 |
| 2exm | ZIP | 1.8 |

Table B.3.117 PDB structures used to construct the CRANkS grids for run 5 of 50 structures for target CDK2

DHFR, 5 structures, Run 1

| PDB Code | Ligand Residue Code | Resolution / Å |
|---|---|---|
| 1u72 | MTX | 1.9 |
| 3ghc | GHC | 1.3 |
| 3ntz | 3TZ | 1.35 |
| 4keb | 1QZ | 1.45 |
| 4qjc | IXE | 1.62 |

Table B.3.118 PDB structures used to construct the CRANkS grids for run 1 of 5 structures for target DHFR

DHFR, 5 structures, Run 2

| PDB Code | Ligand Residue Code | Resolution / Å |
|---|---|---|
| 1mvs | DTM | 1.9 |
| 1u72 | MTX | 1.9 |
| 3noh | TOP | 1.6 |
| 4keb | 1QZ | 1.45 |
| 5hpb | 63W | 1.65 |

Table B.3.119 PDB structures used to construct the CRANkS grids for run 2 of 5 structures for target DHFR

DHFR, 5 structures, Run 3

| PDB Code | Ligand Residue Code | Resolution / Å |
|---|---|---|
| 1dlr | MXA | 2.3 |
| 1hfp | MOT | 2.1 |
| 4qjc | IXE | 1.62 |
| 5ht5 | 65H | 1.9 |
| 5hui | 65N | 1.46 |

Table B.3.120 PDB structures used to construct the CRANkS grids for run 3 of 5 structures for target DHFR

DHFR, 5 structures, Run 4

| PDB Code | Ligand Residue Code | Resolution / Å |
|---|---|---|
| 1u71 | MXA | 2.2 |
| 3eig | MTX | 1.7 |
| 3nxo | D2B | 1.35 |
| 3nxx | D2D | 1.35 |
| 3nxy | D2H | 1.9 |

Table B.3.121 PDB structures used to construct the CRANkS grids for run 4 of 5 structures for target DHFR

DHFR, 5 structures, Run 5

| PDB Code | Ligand Residue Code | Resolution / Å |
|---|---|---|
| 1dls | MTX | 2.3 |
| 1kms | LIH | 1.09 |
| 1ohj | COP | 2.5 |
| 1u71 | MXA | 2.2 |
| 3f8y | DH1 | 1.45 |

Table B.3.122 PDB structures used to construct the CRANkS grids for run 5 of 5 structures for target DHFR

DHFR, 10 structures, Run 1

| PDB Code | Ligand Residue Code | Resolution / Å |
|---|---|---|
| 1u72 | MTX | 1.9 |
| 2c2t | 39B | 1.5 |
| 3f91 | DH1 | 1.9 |
| 3ghc | GHC | 1.3 |
| 3l3r | OAG | 2 |
| 3ntz | 3TZ | 1.35 |
| 3nxt | D2E | 1.7 |
| 4keb | 1QZ | 1.45 |
| 4kfj | 1R0 | 1.76 |
| 4qjc | IXE | 1.62 |

Table B.3.123 PDB structures used to construct the CRANkS grids for run 1 of 10 structures for target DHFR

DHFR, 10 structures, Run 2

| PDB Code | Ligand Residue Code | Resolution / Å |
|---|---|---|
| 1dfr | MTX | 1.7 |
| 1kms | LIH | 1.09 |
| 1mvs | DTM | 1.9 |
| 1u72 | MTX | 1.9 |
| 3gyf | 51P | 1.7 |
| 3noh (second ligand conformation) | TOP | 1.6 |
| 3oaf | OAG | 1.7 |
| 4keb | 1QZ | 1.45 |
| 4m6k | FOL | 1.396 |
| 5hpb | 63W | 1.65 |

Table B.3.124 PDB structures used to construct the CRANkS grids for run 2 of 10 structures for target DHFR

DHFR, 10 Structures, Run 3

| PDB Code | Ligand Residue Code | Resolution / Å |
|---|---|---|
| 1dlr | MXA | 2.3 |
| 1hfp | MOT | 2.1 |
| 2dhf | DZF | 2.3 |
| 3nxt | D2E | 1.7 |
| 3nxx | D2D | 1.35 |
| 3s7a | 684 | 1.8 |
| 4kbn | 25U | 1.84 |
| 4qjc | IXE | 1.62 |
| 5ht5 | 65H | 1.9 |
| 5hui | 65N | 1.46 |

Table B.3.125 PDB structures used to construct the CRANkS grids for run 3 of 10 structures for target DHFR

DHFR, 10 structures, Run 4

| PDB Code | Ligand Residue Code | Resolution / Å |
|---|---|---|
| 1boz | PRD | 2.1 |
| 1dls | MTX | 2.3 |
| 1u71 | MXA | 2.2 |
| 1u72 | MTX | 1.9 |
| 3eig | MTX | 1.7 |
| 3nxo | D2B | 1.35 |
| 3nxx | D2D | 1.35 |
| 3nxy | D2H | 1.9 |
| 4qjc | IXE | 1.62 |
| 5hve | 65Q | 1.46 |

Table B.3.126 PDB structures used to construct the CRANkS grids for run 4 of 10 structures for target DHFR

DHFR, 10 structures, Run 5

| PDB Code | Ligand Residue Code | Resolution / Å |
|---|---|---|
| 1dls | MTX | 2.3 |
| 1kms | LIH | 1.09 |
| 1ohj | COP | 2.5 |
| 1s3w | TQT | 1.9 |
| 1u71 | MXA | 2.2 |
| 3f8y | DH1 | 1.45 |
| 3f91 | DH1 | 1.9 |
| 4keb | 1QZ | 1.45 |
| 4m6l | 21V | 1.7 |
| 5hsu | 63Y | 1.46 |

Table B.3.127 PDB structures used to construct the CRANkS grids for run 5 of 10 structures for target DHFR

DHFR, 25 structures, Run 1

| PDB Code | Ligand Residue Code | Resolution / Å |
|---|---|---|
| 1kms | LIH | 1.09 |
| 1u72 | MTX | 1.9 |
| 2c2s | 34B | 1.4 |
| 2c2t | 39B | 1.5 |
| 2dhf | DZF | 2.3 |
| 3f91 | DH1 | 1.9 |
| 3ghc | GHC | 1.3 |
| 3gi2 | GHW | 1.53 |

| | | |
|---|---|---|
| 3l3r | OAG | 2 |
| 3noh | TOP | 1.6 |
| 3noh (second ligand conformation) | TOP | 1.6 |
| 3ntz | 3TZ | 1.35 |
| 3nxt | D2E | 1.7 |
| 3nxx | D2D | 1.35 |
| 3nxy | D2H | 1.9 |
| 4ddr | MMV | 2.05 |
| 4kbn | 25U | 1.84 |
| 4keb | 1QZ | 1.45 |
| 4kfj | 1Ro | 1.76 |
| 4m6l | 21V | 1.7 |
| 4qjc | IXE | 1.62 |
| 5hqz | 64J | 1.46 |
| 5hsu | 63Y | 1.46 |
| 5hui | 65N | 1.46 |
| 5hvb | 65O | 1.6 |

Table B.3.128 PDB structures used to construct the CRANkS grids for run 1 of 25 structures for target DHFR

DHFR, 25 structures, Run 2

| PDB Code | Ligand Residue Code | Resolution / Å |
|---|---|---|
| 1boz | PRD | 2.1 |
| 1dfr | MTX | 1.7 |
| 1kms | LIH | 1.09 |
| 1mvs | DTM | 1.9 |
| 1pd8 | CO4 | 2.1 |
| 1s3w | TQT | 1.9 |
| 1u71 | MXA | 2.2 |
| 1u72 | MTX | 1.9 |
| 2c2s | 34B | 1.4 |
| 3eig | MTX | 1.7 |
| 3ghc | GHC | 1.3 |
| 3gi2 | GHW | 1.53 |
| 3gyf | 51P | 1.7 |
| 3noh (second ligand conformation) | TOP | 1.6 |
| 3nxo | D2B | 1.35 |
| 3nxr | D2D | 1.35 |
| 3nzd | D2Q | 1.8 |
| 3oaf | OAG | 1.7 |
| 4kak | 06U | 1.8 |
| 4keb | 1QZ | 1.45 |
| 4m6k | FOL | 1.396 |
| 4qjc | IXE | 1.62 |
| 5hpb | 63W | 1.65 |
| 5hqz | 64J | 1.46 |
| 5ht5 | 65H | 1.9 |

Table B.3.129 PDB structures used to construct the CRANkS grids for run 2 of 25 structures for target DHFR

DHFR, 25 Structures, Run 3

| PDB Code | Ligand Residue Code | Resolution / Å |
|---|---|---|
| 1dfr | MTX | 1.7 |
| 1dhf | FOL | 2.3 |
| 1dlr | MXA | 2.3 |
| 1hfp | MOT | 2.1 |
| 1ohj | COP | 2.5 |
| 1yho | TRR | |
| 2dhf | DZF | 2.3 |
| 3eig | MTX | 1.7 |
| 3f8y | DH1 | 1.45 |
| 3ghc | GHC | 1.3 |
| 3gi2 | GHW | 1.53 |
| 3l3r | OAG | 2 |
| 3nuo | 3TU | 1.35 |
| 3nxo | D2B | 1.35 |
| 3nxr | D2D | 1.35 |
| 3nxt | D2E | 1.7 |
| 3nxx | D2D | 1.35 |
| 3s7a | 684 | 1.8 |
| 4kak | 06U | 1.8 |
| 4kbn | 25U | 1.84 |
| 4qjc | IXE | 1.62 |
| 5hqy (second ligand conformation) | 63Y | 1.46 |
| 5ht5 | 65H | 1.9 |
| 5hui | 65N | 1.46 |
| 5hve | 65Q | 1.46 |

Table B.3.130 PDB structures used to construct the CRANkS grids for run 3 of 25 structures for target DHFR

DHFR, 25 structures, Run 4

| PDB Code | Ligand Residue Code | Resolution / Å |
|---|---|---|
| 1boz | PRD | 2.1 |
| 1dhf | FOL | 2.3 |
| 1dls | MTX | 2.3 |
| 1mvt | DTM | 1.8 |
| 1u71 | MXA | 2.2 |
| 1u72 | MTX | 1.9 |
| 2c2t | 39B | 1.5 |
| 3eig | MTX | 1.7 |

| | | |
|---|---|---|
| 3f8y | DH1 | 1.45 |
| 3l3r | OAG | 2 |
| 3nu0 | 3TU | 1.35 |
| 3nxo | D2B | 1.35 |
| 3nxv | D2F | 1.9 |
| 3nxx | D2D | 1.35 |
| 3nxy | D2H | 1.9 |
| 3nzd | D2Q | 1.8 |
| 3s7a | 684 | 1.8 |
| 4kak | 06U | 1.8 |
| 4kd7 | 9DR | 2.715 |
| 4keb | 1QZ | 1.45 |
| 4qjc | IXE | 1.62 |
| 5hqy | 63Y | 1.46 |
| 5hqy (second ligand conformation) | 63Y | 1.46 |
| 5hsu | 63Y | 1.46 |
| 5hve | 65Q | 1.46 |

Table B.3.131 PDB structures used to construct the CRANkS grids for run 4 of 25 structures for target DHFR

DHFR, 25 structures, Run 5

| PDB Code | Ligand Residue Code | Resolution / Å |
|---|---|---|
| 1dls | MTX | 2.3 |
| 1hfp | MOT | 2.1 |
| 1kms | LIH | 1.09 |
| 1ohj | COP | 2.5 |
| 1ohk | COP | 2.5 |
| 1pd8 | CO4 | 2.1 |
| 1s3u | TQD | 2.5 |
| 1s3w | TQT | 1.9 |
| 1u7i | MXA | 2.2 |
| 1yho | TRR | N/A (NMR model, model 1 used) |
| 2dhf | DZF | 2.3 |
| 3f8y | DH1 | 1.45 |
| 3f9i | DH1 | 1.9 |
| 3nu0 | 3TU | 1.35 |
| 3nxx | D2D | 1.35 |
| 3nxy | D2H | 1.9 |
| 3s7a | 684 | 1.8 |
| 4kbn | 25U | 1.84 |
| 4keb | 1QZ | 1.45 |
| 4m6l | 21V | 1.7 |
| 4qhv | IXF | 1.61 |
| 5hpb | 63W | 1.65 |
| 5hqy2 (second ligand conformation) | 63Y | 1.46 |
| 5hsu | 63Y | 1.46 |
| 5ht4 | 65J | 1.6 |

Table B.3.132 PDB structures used to construct the CRANkS grids for run 5 of 25 structures for target DHFR

| DHFR, 50 structures, Run 1<br>PDB Code | Ligand Residue Code | Resolution / Å |
|---|---|---|
| 1boz | PRD | 2.1 |
| 1hfp | MOT | 2.1 |
| 1kms | LIH | 1.09 |
| 1kmv | LII | 1.05 |
| 1mvt | DTM | 1.8 |
| 1pd8 | CO4 | 2.1 |
| 1s3w | TQT | 1.9 |
| 1u71 | MXA | 2.2 |
| 1u72 | MTX | 1.9 |
| 1yho | TRR | N/A (NMR model, model 1 used) |
| 2c2s | 34B | 1.4 |
| 2c2t | 39B | 1.5 |
| 2dhf | DZF | 2.3 |
| 2w3a | TOP | 1.5 |
| 3f8y | DH1 | 1.45 |
| 3f91 | DH1 | 1.9 |
| 3ghc | GHC | 1.3 |
| 3gi2 | GHW | 1.53 |
| 3gyf | 51P | 1.7 |
| 3l3r | OAG | 2 |
| 3noh | TOP | 1.6 |
| 3noh (second ligand conformation) | TOP | 1.6 |
| 3ntz | 3TZ | 1.35 |
| 3nxo | D2B | 1.35 |
| 3nxr | D2D | 1.35 |
| 3nxt | D2E | 1.7 |
| 3nxv | D2F | 1.9 |
| 3nxx | D2D | 1.35 |
| 3nxy | D2H | 1.9 |
| 3oaf | OAG | 1.7 |
| 3s7a | 684 | 1.8 |
| 4ddr | MMV | 2.05 |
| 4g95 | OAG | 1.35 |
| 4kak | 06U | 1.8 |
| 4kbn | 25U | 1.84 |
| 4keb | 1QZ | 1.45 |
| 4kfj | 1R0 | 1.76 |
| 4m6k | FOL | 1.396 |
| 4m6l | 21V | 1.7 |
| 4qhv | IXF | 1.61 |
| 4qjc | IXE | 1.62 |
| 5hpb | 63W | 1.65 |
| 5hqy | 63Y | 1.46 |

| | | |
|---|---|---|
| 5hqz | 64J | 1.46 |
| 5hsu | 63Y | 1.46 |
| 5ht4 | 65J | 1.6 |
| 5ht5 | 65H | 1.9 |
| 5hui | 65N | 1.46 |
| 5hvb | 65O | 1.6 |
| 5hve | 65Q | 1.46 |

Table B.3.133 PDB structures used to construct the CRANkS grids for run 1 of 50 structures for target DHFR

DHFR, 50 structures, Run 2

| PDB Code | Ligand Residue Code | Resolution / Å |
|---|---|---|
| 1boz | PRD | 2.1 |
| 1dfr | MTX | 1.7 |
| 1dhf | FOL | 2.3 |
| 1dlr | MXA | 2.3 |
| 1kms | LIH | 1.09 |
| 1kmv | LII | 1.05 |
| 1mvs | DTM | 1.9 |
| 1ohk | COP | 2.5 |
| 1pd8 | CO4 | 2.1 |
| 1s3w | TQT | 1.9 |
| 1u71 | MXA | 2.2 |
| 1u72 | MTX | 1.9 |
| 1yho | TRR | N/A (NMR model, model 1 used) |
| 2c2s | 34B | 1.4 |
| 2c2t | 39B | 1.5 |
| 2w3a | TOP | 1.5 |
| 3eig | MTX | 1.7 |
| 3f8y | DH1 | 1.45 |
| 3f91 | DH1 | 1.9 |
| 3ghc | GHC | 1.3 |
| 3gi2 | GHW | 1.53 |
| 3gyf | 51P | 1.7 |
| 3noh | TOP | 1.6 |
| 3noh (second ligand conformation) | TOP | 1.6 |
| 3ntz | 3TZ | 1.35 |
| 3nxo | D2B | 1.35 |
| 3nxr | D2D | 1.35 |
| 3nxt | D2E | 1.7 |
| 3nxv | D2F | 1.9 |
| 3nzd | D2Q | 1.8 |
| 3oaf | OAG | 1.7 |
| 3s7a | 684 | 1.8 |
| 4g95 | OAG | 1.35 |
| 4kak | 06U | 1.8 |
| 4kd7 | 9DR | 2.715 |

| | | |
|---|---|---|
| 4keb | 1QZ | 1.45 |
| 4kfj | 1Ro | 1.76 |
| 4m6k | FOL | 1.396 |
| 4m6l | 21V | 1.7 |
| 4qhv | IXF | 1.61 |
| 4qjc | IXE | 1.62 |
| 5hpb | 63W | 1.65 |
| 5hqy | 63Y | 1.46 |
| 5hqy (second ligand conformation) | 63Y | 1.46 |
| 5hqz | 64J | 1.46 |
| 5ht4 | 65J | 1.6 |
| 5ht5 | 65H | 1.9 |
| 5hui | 65N | 1.46 |
| 5hvb | 65O | 1.6 |
| 5hve | 65Q | 1.46 |

Table B.3.134 PDB structures used to construct the CRANkS grids for run 2 of 50 structures for target DHFR

DHFR, 50 structures, Run 3

| PDB Code | Ligand Residue Code | Resolution / Å |
|---|---|---|
| 1dfr | MTX | 1.7 |
| 1dhf | FOL | 2.3 |
| 1dlr | MXA | 2.3 |
| 1dls | MTX | 2.3 |
| 1hfp | MOT | 2.1 |
| 1mvs | DTM | 1.9 |
| 1ohj | COP | 2.5 |
| 1ohk | COP | 2.5 |
| 1pd8 | CO4 | 2.1 |
| 1s3u | TQD | 2.5 |
| 1s3w | TQT | 1.9 |
| 1u71 | MXA | 2.2 |
| 1yho | TRR | |
| 2dhf | DZF | 2.3 |
| 2w3a | TOP | 1.5 |
| 2w3b | FOL | 1.27 |
| 2w3m | FOL | 1.6 |
| 3eig | MTX | 1.7 |
| 3f8y | DH1 | 1.45 |
| 3ghc | GHC | 1.3 |
| 3gi2 | GHW | 1.53 |
| 3gyf | 51P | 1.7 |
| 3l3r | OAG | 2 |
| 3noh | TOP | 1.6 |
| 3noh (second ligand conformation) | TOP | 1.6 |
| 3ntz | 3TZ | 1.35 |
| 3nuo | 3TU | 1.35 |
| 3nxo | D2B | 1.35 |

| | | |
|---|---|---|
| 3nxr | D2D | 1.35 |
| 3nxt | D2E | 1.7 |
| 3nxv | D2F | 1.9 |
| 3nxx | D2D | 1.35 |
| 3nxy | D2H | 1.9 |
| 3oaf | OAG | 1.7 |
| 3s7a | 684 | 1.8 |
| 4g95 | OAG | 1.35 |
| 4kak | 06U | 1.8 |
| 4kbn | 25U | 1.84 |
| 4keb | 1QZ | 1.45 |
| 4m6k | FOL | 1.396 |
| 4qhv | IXF | 1.61 |
| 4qjc | IXE | 1.62 |
| 5hqy (second ligand conformation) | 63Y | 1.46 |
| 5hqz | 64J | 1.46 |
| 5hsu | 63Y | 1.46 |
| 5ht4 | 65J | 1.6 |
| 5ht5 | 65H | 1.9 |
| 5hui | 65N | 1.46 |
| 5hvb | 65O | 1.6 |
| 5hve | 65Q | 1.46 |

Table B.3.135 PDB structures used to construct the CRANkS grids for run 3 of 50 structures for target DHFR

DHFR, 50 structures, Run 4

| PDB Code | Ligand Residue Code | Resolution / Å |
|---|---|---|
| 1boz | PRD | 2.1 |
| 1dhf | FOL | 2.3 |
| 1dls | MTX | 2.3 |
| 1hfp | MOT | 2.1 |
| 1kms | LIH | 1.09 |
| 1mvs | DTM | 1.9 |
| 1mvt | DTM | 1.8 |
| 1ohj | COP | 2.5 |
| 1pd8 | CO4 | 2.1 |
| 1s3u | TQD | 2.5 |
| 1s3w | TQT | 1.9 |
| 1u71 | MXA | 2.2 |
| 1u72 | MTX | 1.9 |
| 2c2s | 34B | 1.4 |
| 2c2t | 39B | 1.5 |
| 2dhf | DZF | 2.3 |
| 2w3a | TOP | 1.5 |
| 2w3b | FOL | 1.27 |
| 3eig | MTX | 1.7 |
| 3f8y | DH1 | 1.45 |

| | | |
|---|---|---|
| 3f91 | DH1 | 1.9 |
| 3gi2 | GHW | 1.53 |
| 3gyf | 51P | 1.7 |
| 3l3r | OAG | 2 |
| 3noh (second ligand conformation) | TOP | 1.6 |
| 3ntz | 3TZ | 1.35 |
| 3nuo | 3TU | 1.35 |
| 3nxo | D2B | 1.35 |
| 3nxr | D2D | 1.35 |
| 3nxt | D2E | 1.7 |
| 3nxv | D2F | 1.9 |
| 3nxx | D2D | 1.35 |
| 3nxy | D2H | 1.9 |
| 3nzd | D2Q | 1.8 |
| 3s7a | 684 | 1.8 |
| 4ddr | MMV | 2.05 |
| 4kak | 06U | 1.8 |
| 4kd7 | 9DR | 2.715 |
| 4keb | 1QZ | 1.45 |
| 4m6l | 21V | 1.7 |
| 4qjc | IXE | 1.62 |
| 5hpb | 63W | 1.65 |
| 5hqy | 63Y | 1.46 |
| 5hqy (second ligand conformation) | 63Y | 1.46 |
| 5hqz | 64J | 1.46 |
| 5hsu | 63Y | 1.46 |
| 5ht4 | 65J | 1.6 |
| 5hui | 65N | 1.46 |
| 5hvb | 65O | 1.6 |
| 5hve | 65Q | 1.46 |

Table B.3.136 PDB structures used to construct the CRANkS grids for run 4 of 50 structures for target DHFR

DHFR, 50 structures, Run 5

| PDB Code | Ligand Residue Code | Resolution / Å |
|---|---|---|
| 1dfr | MTX | 1.7 |
| 1dhf | FOL | 2.3 |
| 1dlr | MXA | 2.3 |
| 1dls | MTX | 2.3 |
| 1hfp | MOT | 2.1 |
| 1kms | LIH | 1.09 |
| 1ohj | COP | 2.5 |
| 1ohk | COP | 2.5 |
| 1pd8 | CO4 | 2.1 |
| 1s3u | TQD | 2.5 |
| 1s3w | TQT | 1.9 |

| | | |
|---|---|---|
| 1u7i | MXA | 2.2 |
| 1yho | TRR | |
| 2c2s | 34B | 1.4 |
| 2c2t | 39B | 1.5 |
| 2dhf | DZF | 2.3 |
| 2w3a | TOP | 1.5 |
| 2w3b | FOL | 1.27 |
| 2w3m | FOL | 1.6 |
| 3f8y | DH1 | 1.45 |
| 3f91 | DH1 | 1.9 |
| 3gi2 | GHW | 1.53 |
| 3gyf | 51P | 1.7 |
| 3l3r | OAG | 2 |
| 3noh (second ligand conformation) | TOP | 1.6 |
| 3ntz | 3TZ | 1.35 |
| 3nuo | 3TU | 1.35 |
| 3nxo | D2B | 1.35 |
| 3nxr | D2D | 1.35 |
| 3nxt | D2E | 1.7 |
| 3nxv | D2F | 1.9 |
| 3nxx | D2D | 1.35 |
| 3nxy | D2H | 1.9 |
| 3nzd | D2Q | 1.8 |
| 3s7a | 684 | 1.8 |
| 4g95 | OAG | 1.35 |
| 4kbn | 25U | 1.84 |
| 4kd7 | 9DR | 2.715 |
| 4keb | 1QZ | 1.45 |
| 4kfj | 1Ro | 1.76 |
| 4m6l | 21V | 1.7 |
| 4qhv | IXF | 1.61 |
| 5hpb | 63W | 1.65 |
| 5hqy | 63Y | 1.46 |
| 5hqy (second ligand conformation) | 63Y | 1.46 |
| 5hqz | 64J | 1.46 |
| 5hsu | 63Y | 1.46 |
| 5ht4 | 65J | 1.6 |
| 5hui | 65N | 1.46 |
| 5hve | 65Q | 1.46 |

Table B.3.137 PDB structures used to construct the CRANkS grids for run 5 of 50 structures for target DHFR

| PDB Code | Ligand Residue Code | Resolution / Å | Reason for Rejection |
|---|---|---|---|
| 1q0b | NAT | 1.9 | N/A |
| 1yrs | L47 | 2.5 | N/A |
| 2fky | N2T | 2.3 | N/A |
| 2fl2 | N4T | 2.5 | N/A |
| 2fl6 | N5T | 2.5 | N/A |
| 2fme | 3QC | 2.1 | N/A |
| 2g1q | N9H | 2.5 | N/A |
| 2gm1 | 2AZ | 2.3 | N/A |
| 2ieh | MOY | 2.7 | N/A |
| 2pg2 | K01 | 1.9 | N/A |
| 2q2y | MKR | 2.5 | N/A |
| 2q2z | MKK | 3.0 | N/A |
| 2uyi | K02 | 2.1 | N/A |
| 2uym | K03 | 2.1 | N/A |
| 2wog | ZZD | 2.0 | N/A |
| 2x7c | KZ9 | 1.9 | N/A |
| 2x7d | EGB | 2.3 | N/A |
| 2x7e | X7E | 2.4 | N/A |
| 2xae | 2XA | 2.6 | N/A |
| 3cjo | K30 | 2.3 | N/A |
| 3k3b | L31 | 2.4 | N/A |
| 3k5e | K5E | 2.0 | N/A |
| 3ken | KEN | 2.5 | N/A |
| 3l9h | EMQ | 2.0 | N/A |
| 4a50 | DQ6 | 2.8 | N/A |
| 4a51 | DQ8 | 2.8 | N/A |
| 4a5y | G7X | 2.5 | N/A |
| 4as7 | 6LX | 2.4 | N/A |
| 4bbg | V02 | 2.8 | N/A |

Table B.4.1 PDB structures for use in the KIF11 dataset.

| PDB Code | Ligand Residue Code | Resolution / Å | Reason for Rejection |
|---|---|---|---|
| 1g5m | N/A | N/A (NMR Structure) | No Ligand |
| 1gjh | N/A | N/A (NMR Structure) | No Ligand |
| 1ysw | 43B | N/A (NMR Structure) | N/A |
| 2021 | 43B | N/A (NMR Structure) | Identical Ligand to 1ysw |
| 2022 | LIU | N/A (NMR Structure) | N/A |
| 202f | LIo | N/A (NMR Structure) | N/A |
| 2w3l | DRO | 2.1 | N/A |
| 4aq3 | 398 | 2.4 | N/A |
| 4ieh | 1.00E+09 | 2.1 | N/A |
| 4lvt | 1XJ | 2.05 | N/A |
| 4lxd | 1XV | 1.9 | N/A |
| 4man | 1Y1 | 2.07 | N/A |
| 5agw | N/A | 2.695 | Ligands between protein structures |
| 5agx | N/A | 2.24 | Ligands between protein structures |
| 5fcg | N/A | 2.1 | No Ligand |
| 5jsn | N/A | 2.1 | No Ligand |

Table B.4.2 PDB structures for use in the BCL2 dataset.

| PDB Code | Ligand Residue Code | Resolution / Å | Reason for Rejection |
|---|---|---|---|
| 3cqu | CQU | 2.2 | N/A |
| 3cqw | CQW | 2.0 | N/A |
| 3mv5 | XFE | 2.5 | N/A |
| 3mvh | WFE | 2.0 | N/A |
| 3ocb | XM1 | 2.7 | N/A |
| 3ow4 | SMY | 2.6 | N/A |
| 3qkk | SMH | 2.3 | N/A |
| 3qkl | SMR | 1.9 | N/A |
| 3qkm | SM9 | 2.2 | N/A |
| 4ekk | ANP | 2.8 | N/A |
| 4ekl | 0RF | 2.0 | N/A |
| 4gv1 | 0XZ | 1.5 | N/A |

Table B.4.3 PDB structures for use in the AKT1 dataset.

| | Grid 1 | Grid 2 | Grid 3 | Grid 4 | Grid 5 |
|---|---|---|---|---|---|
| HSP90 | 1yet | 1uy6 | 2bto | 1uye | 1uyd |
| | 2fwy | 1uy8 | 2qf6 | 1uyk | 2bto |
| | 2xk2 | 2h55 | 3ekr | 2ccu | 2fwy |
| | 3k98 | 2wi6 | 3ft5 | 2vcj | 2wi6 |
| | | 2xk2 | 3ft8 | 3hhu | 3hz5 |
| ADRB2 | 3d4s | 2rh1 | 2rh1 | 2rh1 | 2rh1 |
| | 3ny8 | 3d4s | 3ny8 | 3d4s | 3d4s |
| | 3ny9 | 3ny9 | 3nya | 3ny8 | 3ny8 |
| | 3nya | 3nya | 3p0g | 3ny9 | 3p0g |
| | 3p0g | 3p0g | 3pds | 3nya | 3pds |
| DHFR | 1u72 | 1mvs | 1dlr | 1u71 | 1dls |
| | 3ghc | 1u72 | 1hfp | 3noh | 1kms |
| | 3ntz | 3noh | 4qjc | 3nxo | 1ohj |
| | 4keb | 4keb | 5ht5 | 3nxx | 1u71 |
| | 4qjc | 5hpb | 5hui | 3nxy | 3f8y |
| KIFII | 1q0b | 2m3 | 1q0b | 2gm1 | 2q2y |
| | 3ken | 2uym | 2fme | 2xae | 2wog |
| | 4a5y | 2x7d | 3cjo | 3k5e | 2x7d |
| | 4as7 | 4a5y | 3k3b | 3ken | 3krn |
| | 4bbg | | | 4bbg | 3l9h |
| BCL2 | 1ysw | 1ysw | 1ysw | 2022 | 2022 |
| | 2022 | 202f | 202f | 2w3l | 4aq3 |
| | 2w3l | 2w3l | 4aq3 | 4aq3 | 4ieh |
| | 4ieh | 4ieh | 4ieh | 4lvt | 4lvt |
| | 4lvt | 4man | 4man | 4lx | 4lxd |
| AKT1 | 3cqw | 3cqu | 3qkl | 3cqu | 3cqu |
| | 3mv5 | 3ocb | 3qkm | 3mv5 | 3ocb |
| | 4ekk | 3ow4 | 4ekk | 3ocb | 3ow4 |
| | 4ekl | 4ekk | 4gv1 | 3ow4 | 3qkm |

Table B.4.4 PDB structures used in each grid for testing AutoCRANkS.

| Target | PDB Structure to Dock to | Centre of Grid | Number of X Points | Number of Y points | Number of Z points |
|---|---|---|---|---|---|
| HSP90 | 1uy6 | (2.16, 10.65, 24.07) | 45 | 73 | 61 |
| ADRB2 | 3ny9 | (2.56, 3.61, 52.77) | 17 | 77 | 49 |
| DHFR | 1s3v | (-3.63, 28.98, 4.96) | 57 | 81 | 45 |
| KIF11 | 3k5e | (28.64, 12.55, 15.29) | 51 | 75 | 71 |
| BCL2 | 2w3l | (25.94, 42.59, 5.25) | 41 | 131 | 45 |
| AKT1 | 3qkl | (-8.39, 0.46, 17.66) | 59 | 49 | 43 |

Table B.4.5. Details of the docking parameters required for the PDB structure to dock to for each target. Coordinates are given with respect to the PDB structure to dock to.

# Chapter 5

| SGC Filename | Abbreviated Code | SMILES of Ligand |
|---|---|---|
| NUDT22A-x0196_event1.pdb | x0196 | c1cc(ccc1NC(=O)C[NH+]2CCCC2)F |
| NUDT22A-x0243_event1.pdb | x0243 | C[NH2+]Cc1ccncc1 |
| NUDT22A-x0290_event1.pdb | x0290 | Cn1c(c(cn1)NC(=O)c2ccccc2)C(=O)N |
| NUDT22A-x0329_event1.pdb | x0329 | c1cc(cnc1)C[NH2+][C@H]2CS(=O)(=O)C[C@@H]2O |
| NUDT22A-x0391_event1.pdb | x0391 | c1c(nns1)C(=O)N2CCOCC2 |
| NUDT22A-x0421_event1.pdb | x0421 | c1ccc(cc1)CNc2ccc(cc2)C(=O)[O-] |
| NUDT22A-x0449_event2.pdb | x0449 | c1ccc2c(c1)ncn2Cc3ccc(cc3)F |
| NUDT22A-x0462_event1.pdb | x0462 | c1cc(cnc1)CNc2ccc(cc2N)C(=O)[O-] |
| NUDT22A-x0482_event1.pdb | x0482 | c1cc(oc1)C(=O)NCCc2ccc(cc2)F |
| NUDT22A-x0520_event1.pdb | x0520 | CC(C)(CO)NC(=O)c1cccc(c1)Cl |
| NUDT22A-x0527_event2.pdb | x0527 | CC(=O)N1CCC(CC1)C#N |
| NUDT22A-x0530_event1.pdb | x0530 | Cc1ccc(s1)CNc2cccc(c2)C(=O)[O-] |
| NUDT22A-x0589_event1.pdb | x0589 | c1cc(ccc1NNC(=O)C2CCC2)F |
| NUDT22A-x0703_event1.pdb | x0703 | COc1ccc(cc1)NCc2ccc(cc2)F |
| NUDT22A-x0813_event1.pdb | x0813 | CC1([C@@H](SC(=S)N1C)NO)C |
| NUDT22A-x0826_event1.pdb | x0826 | c1cnccc1CNC(=O)CC#N |

Table B.5.1 Structure files for high-confidence hits determined by the SGC for target NUDT22 as referred to in Section 6.2.2. The abbreviated code for each structure and the SMILES of the ligand in the structure are also shown. Each of the structures can be downloaded from the SGC website (https://www.thesgc.org/ligand-bounds/nudt22-1).