

SuCOS is Better than RMSD for Evaluating Fragment Elaboration and Docking Poses

*Susan H. Leung,^{†§¶} Michael J. Bodkin,[‡] Frank von Delft,[§] Paul E. Brennan,^{§¶} and
Garrett M. Morris^{†*}*

[†]Department of Statistics, University of Oxford, Oxford, OX1 3LB, United Kingdom

[§]Structural Genomics Consortium (SGC) & Target Discovery Institute (TDI), Nuffield
Department of Medicine, University of Oxford, Oxford, OX3 7FZ, United Kingdom

[¶]Diamond Light Source (DLS), Harwell Science and Innovation Campus, Didcot, OX11 0DE,
United Kingdom

[¶]Alzheimer's Research UK, Oxford Drug Discovery Institute and Target Discovery Institute
(TDI), Nuffield Department of Medicine, University of Oxford, Oxford, OX3 7FZ, United
Kingdom

[‡]Evotec, 114 Milton Park, Abingdon, OX14 4RZ, United Kingdom

ABSTRACT

One of the fundamental assumptions of fragment-based drug discovery is that the fragment's binding mode will be conserved upon elaboration into larger compounds. The most common way

of quantifying binding mode similarity is Root Mean Square Deviation (RMSD), but Protein Ligand Interaction Fingerprint (PLIF) similarity and shape-based metrics are sometimes used. We introduce SuCOS, an open-source shape and chemical feature overlap metric. We explore the strengths and weaknesses of RMSD, PLIF similarity, and SuCOS on a dataset of X-ray crystal structures of paired elaborated larger and smaller molecules bound to the same protein. Our redocking and cross-docking studies show that SuCOS is superior to RMSD and PLIF similarity. When redocking, SuCOS produces fewer false positives and false negatives than RMSD and PLIF similarity; and in cross-docking, SuCOS is better at differentiating experimentally-observed binding modes of an elaborated molecule given the pose of its non-elaborated counterpart. Finally we show that SuCOS performs better than AutoDock Vina at differentiating actives from decoy ligands using the DUD-E dataset. SuCOS is available at <https://github.com/susanhleung/SuCOS>.

INTRODUCTION

Fragment Based Drug Discovery (FBDD) is now a well-established method for efficiently exploring chemical space. By experimentally screening a fragment library, initial hit compounds are synthetically elaborated to give molecules with higher molecular weight, greater affinity and selectivity. A variety of computational tools, such as hotspot identification, molecular docking, and molecular dynamics simulations, can help to prioritize which follow-up compounds to synthesize, by predicting information such as binding pose and binding affinities.

In the elaboration of a fragment hit, it is assumed that the binding mode of the original fragment hit in the elaborated fragment molecule will be structurally conserved. Malhotra and Karanicolas recently assembled a dataset of 297 ligand pairs from the PDDBind database to

validate this hypothesis, and found that it was true in 86% of the cases.¹ They quantified the degree of binding mode conservation by devising a combined overlap score, COS, that considers steric overlap and the overlap of chemical features such as hydrogen bond donors and acceptors. COS is an asymmetric metric, as it considers the proportion of overlap of the query molecule with the reference molecule, but not the other way around. COS ranges from 0, where there is no overlap of the elaborated molecule's volume and chemical features onto the non-elaborated counterpart's, to 1, where there is complete overlap. Malhotra and Karanicolas used a COS cutoff that ranged from 0.4 to 0.55, depending on the chemical substructure match between the reference and the query: if a pair had a COS score less than this, the pair was deemed not to have a conserved binding mode.

Drwal *et al.* recently published a related analysis that investigated the similarity of binding modes for fragments compared to even smaller crystallization additives.² They used shape and chemical feature overlap to determine pose similarity, but Protein Ligand Interaction Fingerprints (PLIFs) to determine binding mode similarity. A PLIF similarity threshold of 0.6 or greater was deemed to be a conserved binding mode.

Numerous studies have also shown that the use of binding mode information can lead to greater docking success by selecting the correct binding pose.³⁻⁷ Marcou and Rognan used PLIFs to rescore docking poses and picked poses based on those with the highest PLIF similarity to the PLIF of the reference protein-ligand complex.³ Shape similarity has also been shown to be a successful metric in prioritizing the correct pose.^{5,8}

Despite the existence of these metrics, a root mean square deviation, or RMSD, cutoff of 2.0 Å is still the most widely used measure for assessing whether or not a docking is successful.⁹⁻¹¹ The RMSD is calculated by measuring the positional deviation, or distance, between equivalent

atoms in the reference and query molecules. In fragment-based elaboration, the reference is the fragment hit, and the query is the elaborated fragment. Exactly which atoms are used in the RMSD calculation can vary: sometimes it is a simplistic one-to-one mapping of atoms in the compared molecules, and sometimes it takes chemical symmetry into account, such as the 3-fold rotational equivalence of tertiary-butyls.¹²

If RMSD is calculated between dissimilar molecules, such as fragments and their elaborated counterparts, the common substructure to be compared must be defined. This is commonly done by computing the maximum common substructure, or MCS.^{7,13} Sometimes, a strict match of atoms does not exist, but it is still reasonable to map ring atoms in one molecule to non-ring atoms in another, or atoms in aliphatic rings to those in aromatic rings. Further complications arise when MCS is used for virtual screening libraries of fragment follow-ups, as they can include bioisosteres, pseudo-symmetric small molecules and small changes in the chemical scaffold of that fragment, thus reducing the MCS to less than the original fragment hit. Although manual inspection of compound overlap by experienced structure-based compound designers is effective in qualitatively evaluating compound overlay, these confounding factors make algorithmic comparison of $10^3 - 10^7$ molecules challenging. Nevertheless, RMSD is still widely used for comparing non-identical molecules such as in virtual screening programs.¹⁴⁻¹⁸

Thus, it is unclear which metric is best at quantifying the degree of binding mode conservation, especially in cases where a smaller ligand and its elaborated counterpart are compared. Previous studies have discussed the pitfalls of certain metrics and hence use multiple metrics alongside one another.¹⁹ To the best of our knowledge, no study has focused on the direct comparison of these metrics to quantify conservation of binding mode of elaborated molecules and their non-elaborated counterparts. We utilized the Malhotra and Karanicolas ligand pair dataset to

investigate three metrics: RMSD, PLIF similarity, and a new metric called SuCOS, a combined shape-chemical feature-based metric we have developed. In our study, we differentiate between the calculation of the RMSD between identical molecules, and the RMSD between elaborated molecules with their non-elaborated counterparts. For the former, we shall use the notation ‘All-RMSD’ to indicate that all atoms in the reference and query molecules were used; while for the latter, we use the notation ‘MCS-RMSD’ to indicate that an MCS was first identified in both molecules to define the corresponding pairs of atoms for the RMSD calculation. Furthermore, we have used two measures for PLIF similarity – Tversky and Tanimoto which are denoted TvPLIF and TnPLIF respectively.

Lastly, we investigate the ability of SuCOS to differentiate actives from decoy ligands in the DUD-E dataset²⁰ and compare its performance against the AutoDock Vina predicted affinity.

METHODS

In the current study we used the dataset curated by Malhotra and Karanicolas¹ that consists of 297 ligand pairs from the PDBbind database, where each pair consisted of a smaller and larger ligand solved in complex with the same protein partner. The larger ligand could have, but not necessarily, arisen through synthetic elaborations of the smaller ligand. Here, we perform a direct comparison of the three metrics on (i) Malhotra and Karanicolas’ dataset of paired larger and smaller molecules bound to the same protein (Fig. 1a); (ii) redocking each ligand to its respective protein (Fig. 1b); and (iii) cross-docking of the larger molecule into the smaller molecule’s cognate protein structure (Fig. 1c). We use this dataset to simulate elaboration efforts and situations where virtual screening may have used binding pose similarity to compare virtual molecules to a fragment hit and decide which elaborated molecule to make next. From here on, we shall refer to this dataset as the MK dataset.

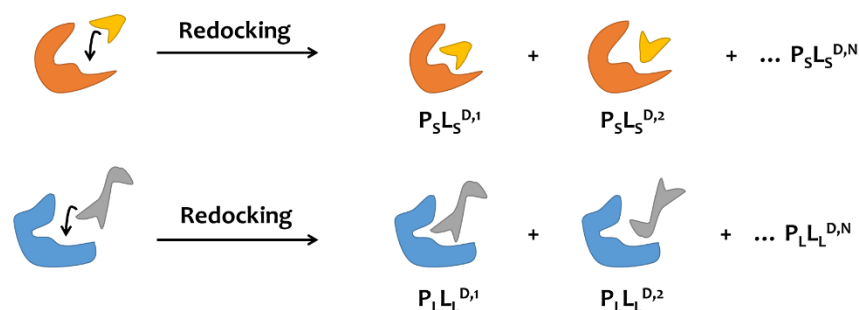
Table 1: Notation used to describe the structures and metrics used in this study.

Notation	Description
L_S^X	Crystal structure of smaller ligand
L_L^X	Crystal structure of larger ligand
L_L^A	Aligned structure of larger ligand, after aligning larger ligand's cognate protein onto the protein from the smaller ligand's complex
P_S^X	Protein in the smaller ligand's crystal structure
P_L^X	Protein in the larger ligand's crystal structure
$P_S L_S^X$	Protein-ligand complex for the smaller ligand's crystal structure
$P_L L_L^X$	Protein-ligand complex for the larger ligand's crystal structure
$P_L L_L^A$	Larger ligand's protein complex that has been aligned onto the smaller ligand's protein complex
$P_S L_S^{D,i}$	The i^{th} docked pose of the smaller ligand into the P_S^X
$P_L L_L^{D,i}$	The i^{th} docked pose of the larger ligand into the P_L^X
$Metric(L_S^X, L_L^A)$	The metric between L_S^X and L_L^A ; e.g. $RMSD(L_S^X, L_L^A)$ is the RMSD between L_S^X and L_L^A .

(a) Part I – Compute MCS-RMSD, TvPLIF and SuCOS between the ligands in the aligned crystal structures, $P_S L_S^X$ and $P_L L_L^A$:



(b) Part II – Compute All-RMSD, TnPLIF and SuCOS for rescoring the redocked smaller ligands $P_S L_S^{D,1}, \dots, P_S L_S^{D,N}$ and redocked larger ligands $P_L L_L^{D,1}, \dots, P_L L_L^{D,N}$:



(c) Part III – Compute All-RMSD/MCS-RMSD, TvPLIF and SuCOS on the cross-docked larger ligand into the smaller ligand's protein structure, $P_S L_L^{D,1}, \dots, P_S L_L^{D,N}$:

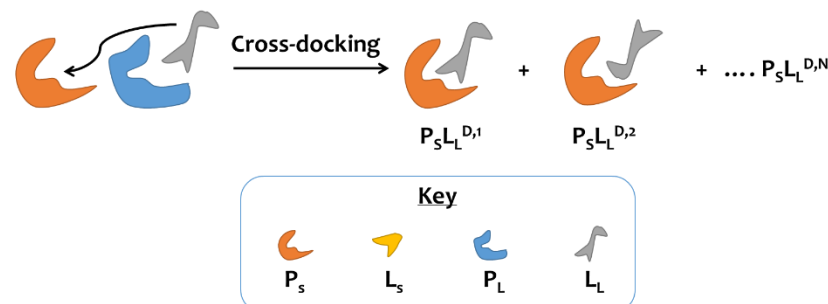


Fig. 1: Overview of study. (a) Computation of metrics, MCS-RMSD, TvPLIF, and SuCOS, to score the conservation of binding mode for the ligands in the aligned crystal structures, $P_S L_S^X$ and $P_L L_L^A$, in the MK dataset. (b) Perform redocking studies to generate poses and score them with the metrics All-RMSD, TnPLIF and SuCOS. (c) Cross-docking the larger ligand, L_L , into the smaller ligand's protein crystal structure, P_S^X , to simulate a realistic virtual screening effort and assessing which metric performs best at picking out the correct pose. In this article we shall denote (a), (b) and (c) with section headers Part I, Part II, Part III respectively.

Downloading and filtering of the Malhotra and Karanicolas ligand pair set

The PDB codes for the ligand pair dataset are available from the supporting information of the Malhotra and Karanicolas study of binding mode changing during chemical elaboration.¹ The dataset contains 297 ligand pairs. However, only 284 ligand pairs were used in this study. Thirteen pairs were excluded, of which 11 pairs have the ligand positioned in between the two chains in at least one of the PDB structures (3vp2/3voz, 3vp4/3uo9, 3voz/3uo9, 3deh/3dek, 3zsy/3zso, 3n7a/3n86, 3zsz/3zso, 3zt1/3zso, 3iaf/3iae, 2pqz/2qnn, 2y54/2y58). Deprecation of the PDB entry 3v0y and absence of 1yp9 from the PDB/UniProt mapping server²¹ led to the exclusion of 3v0y/2qzr and 3ati/1yp9 respectively. The PDBs for the remaining ligand pairs were downloaded from the RCSB PDB Web site²² as the biological assemblies.

Preparation of the Malhotra and Karanicolas ligand pairs

For each PDB structure, only the first chain that contains the small molecule was used. The first model was used for each structure and alternative states were removed using the *removealt* function in PyMOL²³ (version 2.1.0). The larger ligand's crystal structure, $P_L L_L^X$, was aligned to the smaller ligand's crystal structure, $P_S L_S^X$, using PyMOL's *align* function, which we refer to as:

$$P_S L_S^X, P_L L_L^A = align(P_S L_S^X, P_L L_L^X) \quad (1)$$

where $P_L L_L^A$ denotes the larger ligand's structure after alignment. The ligands were chosen by the three letter code according to the Supplementary Information of the study by Malhotra and Karanickolas.¹ For each pair, $P_S L_S^X$ and $P_L L_L^A$, if multiple ligands exist in either structure then only one ligand was used for each. These ligand pairs were chosen by having the shortest distance between their centroids, computed by RDKit's *computecentroid* function. For example, if $P_S L_S^X$ has multiple ligands but $P_L L_L^A$ only has one, then the smaller ligand, L_S , with the smallest distance to the larger, L_L , is chosen. If multiple ligands are present in both, then all pairwise distances between all centroids were calculated and the two closest were chosen.

The smaller and larger PDB structures were checked for consistent numbering of residues between each ligand pair. Residues in pairs with inconsistent numbering were renumbered according to the UniProt numbering scheme using the UniProtKB/SwissProt database.²¹

Protein preparation

The protein was protonated using PDB2PQR²⁴ with the AMBER forcefield and the PROPKA^{25,26} algorithm to assign protonation states at pH 7. The PQR file was then converted into the PDBQT using MGLTools script *prepare_protein4.py* with all default parameters.²⁷

Ligand preparation

The ligands were downloaded without hydrogens in the SDF file format from the RCSB PDB Web site and their respective SMILES obtained from the datafield of the SDF file. These SMILES were then standardized using MolVS 0.0.9²⁸ *standardize_smiles* function. Open Babel 2.4.1²⁹ was used to protonate the SMILES at pH 7. Hydrogens were added to the SMILES using RDKit (2018.09.1) *AddHs* and RDKit *ConstrainedEmbed* was used to generate conformers. The *ConstrainedEmbed* function uses a core molecule as the source of constraint. L_S^X and L_L^X were used as the reference cores for the redocking of L_S and L_L respectively. Hence only one

conformer was generated which corresponds to the protonated crystal pose. For the cross-dockings of the L_L , the reference core was the MCS to L_S^X which was identified by RDKit's *FindMCS* with *completeRingsOnly* and *ringMatchesRingOnly* set to *True*. If there were no amides or aliphatic rings present in the molecule then only one conformer was generated, else ten conformers were generated and clustered, keeping only those which are greater than 0.5 Å from the lowest energy. This was done to create diversity in conformers as the docking treats amides and aliphatic rings as rigid. The conformers were outputted as SDFs that were then converted to MOL2 format using Open Babel in order to retain bond order information. The ligand MOL2 file was then processed into the PDBQT using MGLTools script *prepare_ligand4.py* with all default parameters.²⁷

Docking

In this study, we perform two different types of docking – redocking (Fig. 1b) and cross-docking (Fig. 1c). Redocking is the process of taking a ligand from its structure and docking it back into the same structure. It is typically used before a structure-based virtual screening campaign to validate whether the docking method can successfully recapitulate the experimentally-observed binding pose, and if so, what the best parameters are.

Cross-docking is the process of taking a series of ligand-protein complexes and docking each ligand into every receptor. Cross-docking aims to address the conformational flexibility of the protein to improve docking performance by use of multiple protein structures. In this study, we refer to cross-docking to the specific case of docking the larger ligand into the smaller ligand's protein crystal structure.

AutoDock Vina³⁰ (version 1.2.1) was used for both docking studies. A docking box size of 22 Å on each side was used in all cases. AutoDock Vina generates up to 9 poses for each

docking by default. For the redocking studies, the center was chosen to be the centroid of the crystallographic ligand, either L_S^X or L_L^X depending on the redocking. For the cross-docking study, the box center was set to the centroid of the smaller crystallographic ligand, L_S^X .

For each docking run, AutoDock Vina outputs the docking poses in a PDBQT file which contains the AutoDock Vina predicted affinities (Aff^{Vina}). The PDBQT file was converted into a SDF file using Open Babel that was then used for RMSD and SuCOS calculation. For PLIF calculation, the PDBQT file was converted into a PDB file using Open Babel and combined with either the original PDB of the smaller protein, P_S , or the larger protein, P_L , without the original ligand.

Calculation of RMSD

Both All-RMSD and MCS-RMSD calculations were calculated using RDKit in Python.³¹ For MCS-RMSD, the MCS structure between a reference and query structure was determined by the *FindMCS* function in RDKit, with both options *completeRingsOnly* and *ringMatchesRingOnly* set to *True*. The RMSD calculation takes into account symmetry if present in a molecule such as a para-substituted phenyl ring. However, it does not take into account multiple substructure matches, for example if there are multiple MCSs present in a molecule, it will only match one of them.

Calculation of Protein Ligand Interaction Fingerprints (PLIFs)

PLIFs have been previously reported in various forms, using different definitions for their interactions.^{3,13,32,33} All aim to interpret 3D protein-ligand interactions into a 1D bit array that represent the presence or absence of specific interactions types with specific residues.

In this study, PLIFs were calculated using Arpeggio.³⁴ PDB files were preprocessed using the *clean_pdb.py* Python script from <https://github.com/harryjubb/pdbtools>. We include 12 of

Arpeggio's 15 interactions types: covalent, VdW, hydrogen bond, halogen bond, ionic, aromatic, hydrophobic, carbonyl, polar, metal, weak hydrogen bond and weak polar. Proximal interactions were excluded because they make up a significant proportion (37%) of the total interactions seen in the MK dataset (Supplementary S1). Steric clash and VdW clash interaction types were also excluded. Interactions with water were not considered. If multiple interactions of the same type were made to a specific protein residue then it was only recorded once. Only interactions between the ligand and the protein residues and between the ligand and any metal ions were considered.

We have used two measures for PLIF similarity – Tversky and Tanimoto which are denoted TvPLIF and TnPLIF respectively. The RDKit functions *DataStructs.TanimotoSimilarity* and *DataStructs.TverskySimilarity* were used to calculate the Tanimoto and Tversky coefficients respectively. TvPLIFs were calculated with weights $\alpha = 1$ and $\beta = 0$ corresponding to the smaller and larger ligand respectively or in the case of the cross-docking, for the crystal pose, L_L^X , and docked pose, L_L^D , respectively.

Calculation of SuCOS

SuCOS is a metric inspired by Malhotra and Karanicolas' COS score.¹ SuCOS is composed of half shape overlap and half chemical feature overlap. It utilizes two RDKit functions, *ShapeProtrudeDist* for shape overlap and *ScoreFeats* for chemical feature overlap:

$$SuCOS = 0.5 (1 - ShapeProtrudeDist) + 0.5 (ScoreFeats) \quad (2)$$

The proportion of the reference molecule's volume that is covered by the query molecule is calculated by $(1 - ShapeProtrudeDist)$ with the option *allowReordering* set to *False*. This is an asymmetric shape overlap metric, so if the query molecule completely covers the reference molecule in volume, then the score from $(1 - ShapeProtrudeDist)$ will be 1 regardless of how

much larger the reference molecule is compared to the query molecule. The score from $(1 - \text{ShapeProtrudeDist})$ ranges from 0 (no volume overlap) to 1 (complete volume overlap).

The chemical feature overlap score was calculated using all of RDKit's pharmacophore feature types to create the feature maps before being scored using *ScoreFeats*. It is an asymmetric score as it involves normalization to the smaller ligand. The feature map scoring mode was set to the default *All*, which can output scores greater than 1; therefore, we limited the outputted feature map score to a maximum of 1, since the region of interest is the classification boundary and not at this upper end of the range. As both the shape and chemical feature components of SuCOS are asymmetric, SuCOS is thus asymmetric, *i.e.* it depends only on how well the query overlaps with the reference, and, unlike RMSD, is independent of any size differences between the two.

In Part I, we compare the structure of each larger ligand (reference) to that of its corresponding smaller ligand (query). In Part II and Part III, we compare a docked pose to a reference crystal pose, and for this, the reference molecule and query molecule was the crystal pose and docking pose respectively.

Using SuCOS for Virtual Screening

The DUD-E dataset²⁰ was used to assess the ability of SuCOS to discriminate active from decoy molecules. We used the publically-available predicted binding modes of the actives and decoys,^{35–37} which were generated using the Smina fork³⁸ of AutoDock Vina³⁰. Some molecules could not be parsed by RDKit (see Table S1). SuCOS was calculated for all docked poses of the actives and decoys with respect to the crystallographic ligand of the corresponding DUD-E target. A single SuCOS value was kept, which was the maximum achieved out of all the docked poses for each ligand; similarly for the best AutoDock Vina score. The performance of ranking by SuCOS was compared with ranking by the AutoDock Vina score using the area under the

curve (AUC) of the receiver operating characteristics (ROC) curves, calculated for each of the 102 DUD-E targets. To quantify early enrichment, the ROC enrichment^{39,40}, RE, at 0.5%, 1%, 2%, 5%, was also calculated.

RESULTS AND DISCUSSION

Part I. Comparison of MCS-RMSD, TvPLIF and SuCOS between the ligands in the aligned crystal structures of the Malhotra and Karanicolas ligand pair set.

Malhotra and Karanicolas evaluated the ligand overlap of the smaller and larger ligands by computing the COS score. To compare against this combined overlap metric, we also calculated the MCS-RMSD and TvPLIF for each pair of this dataset (Fig. 1a).

The COS metric uses commercial software ROCS⁴¹ to compute the overlap of volume and chemical features. Volume overlap is the intersection volume of the two molecules, normalized by the smaller molecule. Chemical feature overlap refers to the spatial overlap of pharmacophoric features such as hydrogen bond donors, hydrogen acceptors, aromatic groups *etc.* and likewise it is computed by taking the intersection of the chemical features of the two molecules and normalizing by the chemical features of the smaller.

We devised an open-source alternative to COS, namely SuCOS, which uses RDKit functions *ShapeProtrudeDist* and *ScoreFeats* to compute the volume and chemical feature overlap respectively (see Methods). We compared the performance of SuCOS to COS by computing SuCOS for each pair of the filtered Malhotra and Karanicolas ligand pair set and compared it to the corresponding COS score. SuCOS achieved a good correlation of $R_P = 0.93$ (Fig. 2a). The vertical dotted lines show the three different COS cutoffs Malhotra and Karanicolas used to determine whether an elaborated molecule had changed binding mode. The different cutoffs accounted for pairs with differing chemical substructure scores. Upon inspection of the outliers,

four of the points have low COS but high SuCOS (PDB IDs of smaller/larger pairs: 2hdq/112s, 3adt/3ads, 4e49/3f8e and 3adu/3ads, shown by the purple stars in Fig. 2a); all these structures have multiple smaller and/or larger ligands bound to the protein and it is possible that COS was computed for the one with poorer overlap. It is also possible that another ligand pair (PDB IDs: 1o6i/1w1p) has the smaller and larger ligand swapped (shown by green cross in Fig. 2a) as the reported volume overlap score is 1.06 but the ligand corresponding to 1o6i is the larger of the two.

As both metrics rely on pre-aligned ligands, slight differences in COS and SuCOS values can be attributed to the different methods used to align the protein pairs. Fig. 2b and in Fig. 2c shows the two ligand pairs with the largest differences in COS and SuCOS which can be explained by the algorithmic differences in RDKit functions versus ROCS, such as how each chemical feature point is represented.

From here on, we used SuCOS to calculate the combined shape and chemical feature overlap.

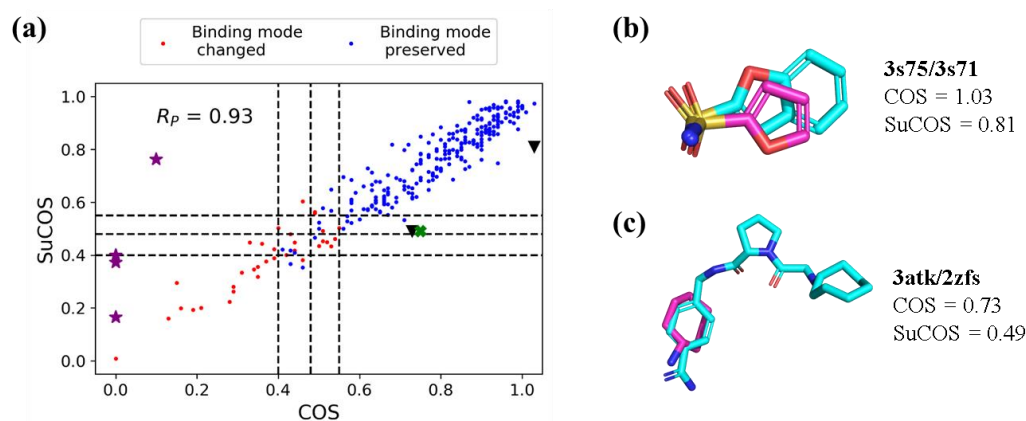


Fig. 2 SuCOS is a good open-source alternative to the COS metric used by Malhotra and Karanicolas. (a) SuCOS was calculated for every ligand pair in the filtered Malhotra and Karanicolas set. The COS values for each ligand pair were obtained from the Supporting Information of the Malhotra and Karanicolas study. Good correlation ($R_p = 0.93$) shows that

SuCOS is an excellent open-source alternative and one that we shall use in this study. Dotted lines show the three different COS cutoff levels Malhotra and Karanickolas used to define the conservation of binding mode. The corresponding three cutoff levels are also shown on the SuCOS axis. As defined by Malhotra and Karanickolas, the conserved binding modes are shown by the blue points, while the unconserved modes are shown by the red points. The four outliers in purple stars and green cross are discussed in the text. The two black triangles represent pairs 3s75/3s71 and 3atk/2zfs, which are the pairs with largest differences in COS and SuCOS values. These are also discussed in the text and are shown in (b) and (c) respectively, where the smaller and larger ligand is shown with magenta and cyan carbons respectively.

Next, to determine the correlation between MCS-RMSD and SuCOS, the two metrics were calculated on each of the ligand pairs. The scatter plot shows the expected trend that MCS-RMSD decreases with SuCOS score, with a Pearson correlation coefficient of -0.66 (Fig. 3a). The widely used 2 Å RMSD cutoff⁴² is also shown as a dotted line. Points located in the top-right and bottom-left can be regarded as false negatives (FNs) (high RMSD, high SuCOS) and false positives (FPs) (low MCS-RMSD, low SuCOS) respectively and the proportion of these points are 12% and 2% respectively. Visual inspection of the ligand pairs corresponding to the FNs gives rise to some of the examples given in Table 2. These cases highlight the pitfalls of using atom-to-atom matching to compute RMSD: if the molecules have pseudosymmetry, multiple substructure matches or substructures which have similar chemical properties, a non-substructure matching alternative such as SuCOS may be more appropriate to use when comparing poses of elaborated molecules against their non-elaborated counterparts.

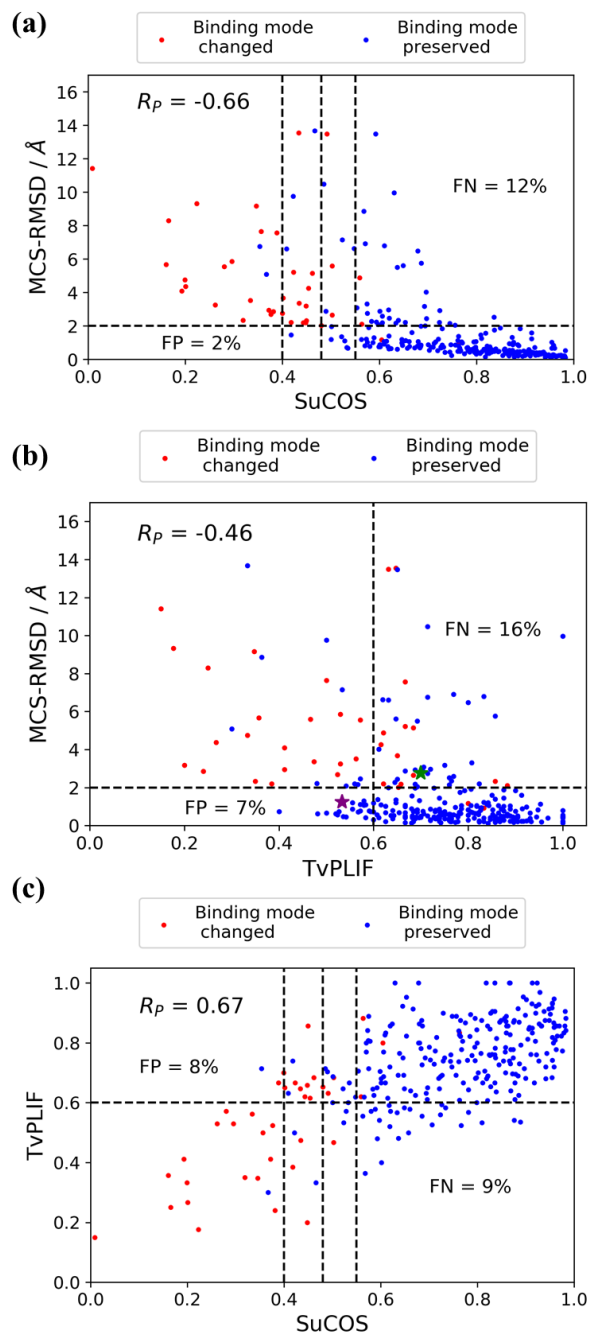
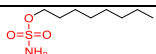

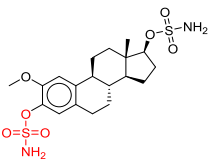
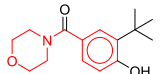
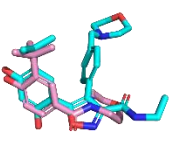
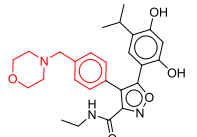
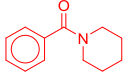
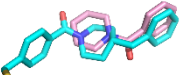
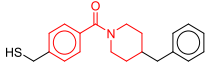
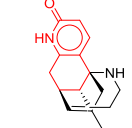
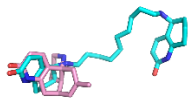
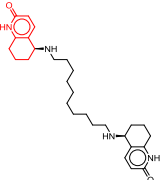


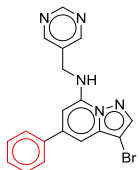


Fig. 3. Comparison of three conservation of binding mode metrics on the Malhotra and Karanicolas ligand pair set. Dotted lines show the commonly used cutoffs to define the conservation of binding mode: MCS-RMSD = 2 Å, TvPLIF = 0.6, COS = 0.55, 0.48, 0.40. (a) Scatter plot of MCS-RMSD against SuCOS. Points located in the upper-right part are considered

as false negative (FN) points ($\text{MCS-RMSD} > 2 \text{ \AA}$ and $\text{SuCOS} > 0.55$), and points located in the lower-left part are considered as false positive (FP) points ($\text{MCS-RMSD} < 2 \text{ \AA}$ and $\text{SuCOS} < 0.55$). (b) Scatter plot of MCS-RMSD against TvPLIF. Points located in the upper-right part are considered as FNs ($\text{MCS-RMSD} > 2 \text{ \AA}$ and $\text{TvPLIF} > 0.6$), and points located in the lower-left part are considered as FPs ($\text{MCS-RMSD} < 2 \text{ \AA}$ and $\text{TvPLIF} < 0.6$). The green and purple star represents the example FN given in Supplementary S2 and example FP given in Supplementary S3 respectively. (c) Scatter plot of TvPLIF against SuCOS. Points located in the lower-right part are considered as FNs ($\text{TvPLIF} < 0.6$ and $\text{SuCOS} > 0.55$), and points located in the upper-left are considered as FPs ($\text{TvPLIF} > 0.6$ and $\text{SuCOS} < 0.55$).

Table 2: Examples of cases where maximum common substructure RMSD is inappropriate for comparing poses of elaborated molecules with fragment hits.^a

PDB ID	2D Ligand representation	3D Representation	Comment
Smaller: 3ibi			$\text{MCS-RMSD} = 13.5 \text{ \AA}$ $\text{TvPLIF} = 0.65$ $\text{SuCOS} = 0.59$ MCS does not compare aliphatic chain against ring system.
Larger: 2gd8			
Smaller: 2xht			$\text{MCS-RMSD} = 6.5 \text{ \AA}$ $\text{TvPLIF} = 0.80$ $\text{SuCOS} = 0.68$ Good overlap of chemical groups with similar chemical properties (bioisostere) but MCS-RMSD calculation compares non-overlapping substructure due to better substructure match.
Larger: 2vci			
Smaller: 4eh4			$\text{MCS-RMSD} = 6.9 \text{ \AA}$ $\text{TvPLIF} = 0.77$ $\text{SuCOS} = 0.57$ Psuedosymmetric molecule but good overlap of alicyclic and aromatic rings.
Larger: 3iw7			

Smaller: 1gpn*			MCS-RMSD = 13.7 Å	Multiple substructure matches in larger molecule
Larger: 1h22*			TvPLIF = 0.33 SuCOS = 0.47	
Smaller: 2vta*			MCS-RMSD = 5.2 Å	Elaboration involves heteroatom changes to core. MCS would result in comparing the benzene ring
Larger: 2r3k*			TvPLIF = 0.68 SuCOS = 0.46	

*These pairs are not FNs but nevertheless represent cases where the RMSD is misleading.

The PDB IDs of the smaller and larger structures are shown next to their corresponding 2D structures, with the common substructures used to compute the MCS-RMSD highlighted in red. The 3D representation shows a protein-based overlay of the fragment hit and larger ligand's crystal structures, generated by aligning the larger ligand's protein structure to that of the smaller ligand using the *align* function in PyMOL.²³ The MCS-RMSD, TvPLIF and SuCOS values for the aligned smaller and larger ligands are also shown, as well as an explanation of why MCS--RMSD is inappropriate.

Plotting MCS-RMSD against TvPLIF shows a weak negative correlation (Fig. 3b, $R_P = -0.46$). The vertical line at TvPLIF = 0.6 is the proportion of interactions that Marcou and Rognan found that must be maintained to correspond to the 2 Å RMSD cutoff.⁴³ Again, several points are situated in the top-right corner and bottom-left corner. These correspond to the FNs and FPs respectively and make up 16% and 7% of the points respectively. 59% of the FNs in Fig. 3b are also FNs in Fig. 3a. Manual inspection of the remaining FNs show indeed a changed binding mode but the ligand manages to retain many of the original interactions. For example in pair 3adt/3ads, the larger ligand has changed binding mode but 70% of interactions are retained (Supplementary S2). Through processing of the Arpeggio output, only the interaction type and

the residue number are kept, but no information is kept about which atom of the ligand is responsible for the interaction. Hence by this definition, the same interaction can be maintained with a different ligand atom which can accommodate some movement in the ligand.

Examination of the FP cases shows indeed very similar binding modes of the pairs, however the lower than expected TvPLIF may be explained by movement in the protein binding pocket and the strict definitions of distance and geometry for a particular interaction. Using 1ce5/1g3c as an example, there is good overlap of the smaller and larger ligand (MCS-RMSD = 1.2 Å, TvPLIF = 0.53, SuCOS = 0.81), however due to conformational change in the protein binding site and slight differences in distances and angles, the TvPLIF is lower than expected (Supplementary S3). Furthermore, many of the interactions present in the smaller complex, but absent from the larger complex, are weak interactions. When computing PLIF similarity, all interaction types are given equal importance, but this can give a lower than expected TvPLIF, as seen in this example and many other FPs.

The plot of TvPLIF against SuCOS shows a slightly better correlation (Fig. 3c, $R_p = 0.67$). The points located in the top-left (SuCOS < 0.55, TvPLIF > 0.6) and bottom-right (SuCOS > 0.55, TvPLIF < 0.6) of the plot are labelled as FPs and FNs, which make up 8% and 9% of the points respectively. Manual inspection of some of the FNs can explain the lower than expected TvPLIFs. For example, in three of the structures missing residues were found in the PDB file (3uok, 3hoz, 1t48). Unlike RMSD and SuCOS which are ligand-centric metrics, PLIF is also dependent on protein structure, hence care must be taken that the quality of the protein structure is adequate.

Furthermore, as the PLIF depends on the conformation of the protein binding side residues, there should be some noise that is a result of protein conformation differences between ligand

pairs in Fig. 3b and Fig. 3c. To investigate this variable, the larger ligands of the aligned structure, L_L^A , were combined with the smaller proteins structure, P_S^X , and the PLIFs computed on $P_S^X L_L^A$. For Fig. 3b and Fig. 3c, R_P was improved to -0.53 and 0.71 respectively (Supplementary S4). Indeed PLIF similarity is able to capture information about which interactions are kept or lost across multiple crystal structures of ligands bound to the same protein that ligand centric metrics such as RMSD and SuCOS cannot. However, if only one protein conformation is used *e.g.* in the redocking or docking numerous virtual ligands into the same protein, then information regarding the ligand pose should be captured using a ligand centric metric.

Part II. Using All-RMSD, TnPLIF and SuCOS to rescore the redockings of the Malhotra and Karanicolas ligand pair set.

Typically during a docking campaign, multiple poses are outputted for each docking run and the pose with the best docking score is usually chosen, although stochastic methods can produce several poses for consideration. RMSD is frequently the default metric for evaluating pose prediction, and the 2 Å cutoff is still widely used: any pose within 2 Å of the crystal ligand pose is deemed a ‘successful’ docking. As shape overlap and PLIF similarity have also been used to measure the conservation of binding mode, we also investigated their use in evaluation of docking success, including computing the rates of FPs and of FNs.

In Part I, an asymmetric Tversky coefficient of shared protein-ligand interactions, TvPLIF, was used to compare binding modes of elaborated molecules against their smaller, non-elaborated counterparts. This was done to prioritize the interactions known to be made by the smaller ligand. However, the Tanimoto coefficient, TnPLIF, of shared interactions has been used in numerous studies to evaluate redocking.^{44,45} TnPLIF measures the similarity between two poses of the same molecule, typically its docked pose and its crystal structure.

In this study, each of the smaller ligands, L_S , was redocked into its cognate protein crystal structure, P_S^X , which produced a number of docked poses, $P_S L_S^{D,i}$, where i is the i^{th} pose produced for that docking. Similarly, the larger ligands, L_L , were redocked into their own cognate protein crystal structures, P_L^X :

$$P_S L_S^{D,i} = Dock(L_S, P_S^X) \quad (3)$$

$$P_L L_L^{D,i} = Dock(L_L, P_L^X) \quad (4)$$

where i is the i^{th} docked pose.

As the MK dataset was uniquified only in terms of PDB pairs and not by single PDB IDs, we extracted the set of unique PDB entries in the MK dataset, to give 485 unique PDB IDs for redocking using AutoDock Vina. From this, 436 ligands were successfully redocked into their cognate proteins. The remaining 49 failed because of various errors, including the protein crystal structure having incomplete loops; the ligand having unusual atom types, such as boron; or the docked ligand SDF being unreadable by RDKit. For each ligand, up to 9 poses were generated (see Methods). In total, 3793 poses were produced and for each pose All-RMSD, TnPLIF and SuCOS were computed with reference to its crystal structure pose. Each pose also had an associated Vina score, Aff^{Vina} . Each metric was then used to rank the poses and select a single pose, *i.e.* the pose with the lowest RMSD, $Pose(RMSD_{best})$; that with the highest TnPLIF, $Pose(TnPLIF_{best})$; that with the highest SuCOS, $Pose(SuCOS_{best})$; and that with the lowest Aff^{Vina} , $Pose(Aff_{best}^{Vina})$:

$$Pose(RMSD_{best}) = Min(RMSD(L_S^{D,1}, L_S^X), RMSD(L_S^{D,2}, L_S^X), \dots RMSD(L_S^{D,N}, L_S^X)) \quad (5)$$

$$Pose(TnPLIF_{best}) = Max(TnPLIF(L_S^{D,1}, L_S^X), TnPLIF(L_S^{D,2}, L_S^X), \dots TnPLIF(L_S^{D,N}, L_S^X)) \quad (6)$$

$$Pose(SuCOS_{best}) = Max(SuCOS(L_S^{D,1}, L_S^X), SuCOS(L_S^{D,2}, L_S^X), \dots SuCOS(L_S^{D,N}, L_S^X)) \quad (7)$$

$$Pose(Aff_{best}^{Vina}) = Min(Aff^{Vina}(L_S^{D,1}), Aff^{Vina}(L_S^{D,2}), \dots Aff^{Vina}(L_S^{D,N})) \quad (8)$$

where N is the total number of docked poses produced for that ligand. If multiple poses had the same best score, then only the first occurring pose was kept for that metric.

We looked at the effect of varying the cutoff for All-RMSD, TnPLIF, and SuCOS on the proportion of dockings that matched the crystallographic binding mode (Supplementary Fig. S5). Using the standard 2 Å cutoff, 358 of the 436 redockings (82%) generated at least one ‘successful’ pose. Ranking the docked poses using the other metric, TnPLIF, SuCOS, and Aff^{Vina} , we found 315 (72%), 331 (76%), and 246 (56%) of the poses were within 2 Å of the crystallographic binding mode, respectively (Supplementary Fig. S5a). It is worth noting that the success rate when ranking docked poses by the native Vina score, 56%, was less than the reported 78% success rate.³⁰

In addition to All-RMSD, the success of pose prediction can be defined by SuCOS or TnPLIF (Supplementary Fig. S5 (b) and (c) respectively). Ranking poses by Aff^{Vina} consistently performed the worst at pose prediction. The difficulty of accurate affinity prediction has already been well established and is supported by the recent study of Ramírez and Caballero.⁴⁶ Furthermore, the standard error of the AutoDock Vina scoring function is 2-3 kcal/mol,²⁷ so if there is little variation between the affinity of the poses, it will perform poorly in picking the ‘correct’ pose. Rescoring by TnPLIFs and SuCOS leads to greater pose prediction success as seen in previous studies.^{2,4-6,8,43}

All-RMSD and SuCOS performed similarly across all three methods of defining success, which suggests that SuCOS is a good non-substructure matching alternative to All-RMSD. The 2 Å All-RMSD cutoff corresponds to approximately SuCOS = 0.55 and TnPLIF = 0.4 (Supplementary Fig S5). This TnPLIF value is lower than the 0.6 TnPLIF cutoff found by Marcou and Rognan.³ However this may be due to differences in the datasets, and/or due to the

difference in the number and types of interactions used to generate the PLIF. For example, Arpeggio has polar and weak polar interaction types, whereas Marcou and Rognan's in-house method did not.

Where TnPLIF disagrees with all other metrics, it is again worth noting that PLIFs are highly sensitive to distance and direction. Taking the redocking of ligand of 1o39 as an example (Fig. 4), another disadvantage of ranking by TnPLIF is made clear as penalties are introduced when a docked pose makes more interactions than the crystallographic binding mode. Furthermore, the ligand binds to the surface of the protein and is partially solvent exposed. This part of the ligand that does not interact with any residues is not captured in the PLIF. Unlike RMSD and SuCOS, the nature of the binding site can influence how much of the ligand pose is captured.

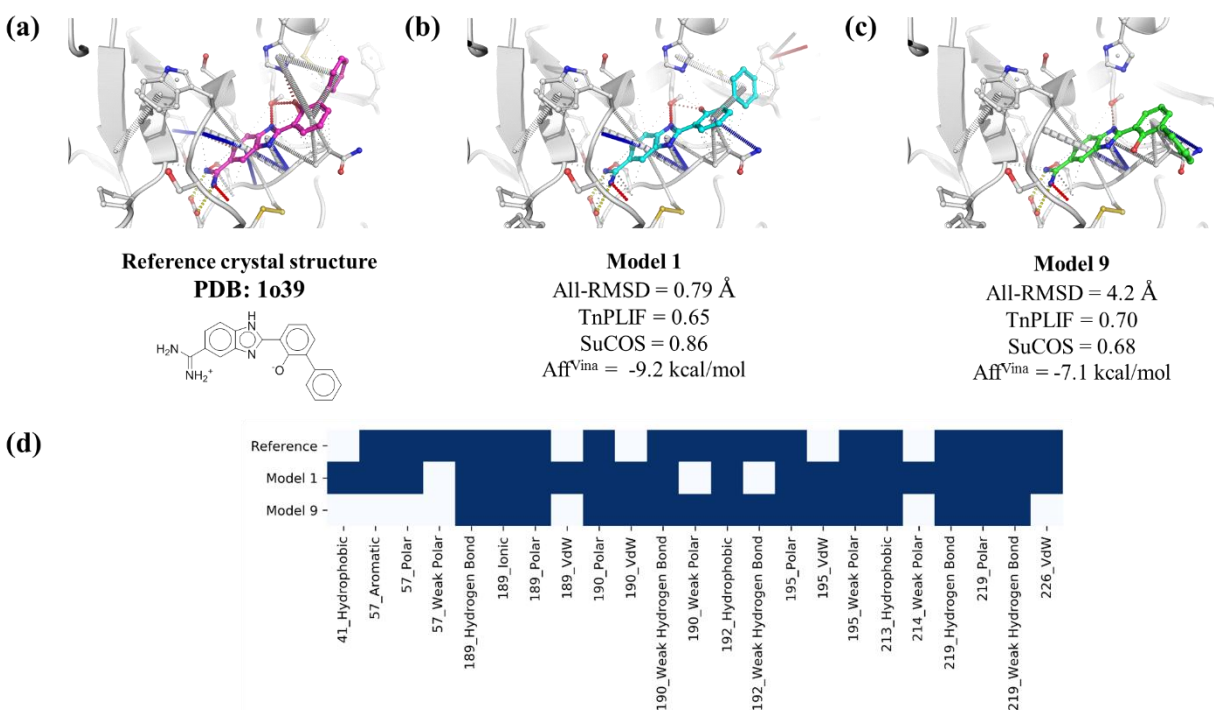


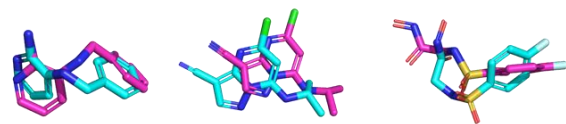
Fig. 4 Redocking ligand of 1o39: example of a disadvantage of using TnPLIF when redocking. Both model 1 and model 9 have very similar TnPLIF scores; however, model 1 is very similar to the crystal structure pose and model 9 is quite different. (a) The crystallographic binding mode of

the ligand is shown in pink sticks. (b) Model 1 is shown in cyan sticks and is ranked best by All-RMSD, SuCOS and Aff^{Vina} . The pose overlaps almost exactly with the crystal structure of the ligand but it forms several additional interactions by just slight differences in orientation of the ligand. (c) Model 9 is shown in green sticks and is ranked best by TnPLIF. It makes fewer interactions than are present in the crystal structure, but the terminal phenyl ring is incorrectly pointing out of the binding pocket. (d) PLIF interaction heatmap of the reference, Model 1 and Model 9. The presence of an interaction is shown by a blue square. The notation for the PLIFs along the x -axis is the residue number and the interaction type: *e.g.* 41_Hydrophobic refers to a hydrophobic interaction with residue 41.

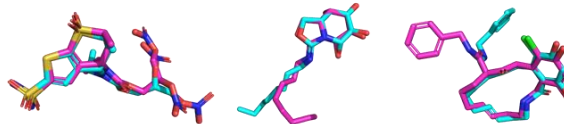
Table 3: Visual inspection of all the poses from the redockings shows that all metrics – All-RMSD, TnPLIF and SuCOS – give false positives and false negatives.^a

Criterion	TP	TN	FP	FN
All-RMSD < 2 Å	336	70	22	8
TnPLIF > 0.60	211	91	1	133
SuCOS > 0.55	338	88	4	6

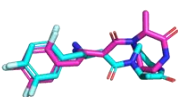
^aThe table gives the numbers of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) for each metric and criterion.

(a) All-RMSD FPs

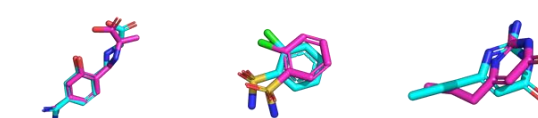
2ohm	2vtr	3f19
RMSD = 1.5 Å	RMSD = 2.0 Å	RMSD = 1.9 Å
TnPLIF = 0.58	TnPLIF = 0.30	TnPLIF = 0.33
SuCOS = 0.52	SuCOS = 0.37	SuCOS = 0.48

(b) All-RMSD FNs

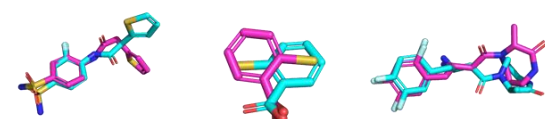
3k2f	2wc3	2xx5
RMSD = 3.1 Å	RMSD = 2.7 Å	RMSD = 3.5 Å
TnPLIF = 0.60	TnPLIF = 0.59	TnPLIF = 0.52
SuCOS = 0.69	SuCOS = 0.67	SuCOS = 0.67

(c) TnPLIF FPs

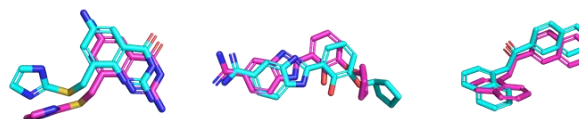
2iiv
RMSD = 2.2 Å
TnPLIF = 0.68
SuCOS = 0.67

(d) TnPLIF FNs

1g3c	2weh	3hvg
RMSD = 1.0 Å	RMSD = 1.5 Å	RMSD = 1.4 Å
TnPLIF = 0.33	TnPLIF = 0.39	TnPLIF = 0.39
SuCOS = 0.64	SuCOS = 0.58	SuCOS = 0.61

(e) SuCOS FPs

3r17	4e4a	2iiv
RMSD = 2.5 Å	RMSD = 2.8 Å	RMSD = 2.2 Å
TnPLIF = 0.49	TnPLIF = 0.34	TnPLIF = 0.68
SuCOS = 0.55	SuCOS = 0.56	SuCOS = 0.67

(f) SuCOS FNs

1k4g	1o3p	1kyn
RMSD = 1.5 Å	RMSD = 1.8 Å	RMSD = 1.3 Å
TnPLIF = 0.55	TnPLIF = 0.44	TnPLIF = 0.29
SuCOS = 0.52	SuCOS = 0.52	SuCOS = 0.54

Fig. 5 (a) – (f) Three examples of FPs and FNs for each metric are shown. The crystal ligand is shown in cyan and the docked pose in pink throughout. The PDB code is shown under each structure, together with the All-RMSD, TnPLIF and SuCOS values for the docked pose against the crystal ligand pose. The values highlighted in red are the false values.

In the case of redocking, the reference and query molecule are identical, so All-RMSD is used. When comparing molecules and their elaborated counterparts, RMSD depends on defining the pairs of corresponding atoms to be used in the calculation of the positional deviations. Some of the problems shown in the examples where RMSD is inappropriate to use for elaborated molecules in Part I (Table 2) may also arise when comparing identical molecules. Therefore, all

poses for all redockings were visually inspected and each redocking was manually classified as “successful” (had at least one pose that closely matched the crystallographic binding mode), or “unsuccessful” (had no poses that closely matched the crystallographic binding mode). This classification is indeed somewhat subjective, however, as each metric may contain FPs and FNs, it was done so that the strengths and weaknesses of metric can be understood.

Thus, for each metric the number of TPs (*i.e.*, the metric correctly classified a successful docking), TNs (*i.e.*, the metric correctly classified an unsuccessful docking), FPs (*i.e.*, incorrectly classified a successful docking) and FNs (*i.e.*, incorrectly classified an unsuccessful docking) was recorded (Table 3).

Visual inspection of the FPs and FNs for All-RMSD again highlighted some of its weaknesses. For example, there were numerous cases of FPs where the molecule is small. The RMSD metric is size dependent^{47,48} and using the 2 Å cutoff for smaller ligands may be too large. Normalization of RMSD by molecular size, or using size-appropriate cutoffs for different sized molecules may overcome this, yet the 2 Å cutoff is still widely used. For All-RMSD FNs, there were cases where there is good overlap of the cores of the crystal ligand and docked pose but the side chain has changed conformation (see 2wc3 and 2xx5 in Fig. 5b). This dramatically increases the All-RMSD value, as discussed by Hawking *et al.*⁴⁷ but SuCOS is much less affected. Pseudosymmetry led to a high All-RMSD for 3k2f despite the good overlap of the docked and crystal ligand pose.

The relatively low number of TPs and high number of FNs for TnPLIF can be attributed to the cutoff of TnPLIF = 0.6 being stricter and not equivalent to the 2 Å RMSD cutoff as discussed before. Indeed, reducing this cutoff to TnPLIF = 0.4 decreases the number of FNs and FPs to 7 and 37 respectively.

Several of the SuCOS FPs resemble the All-RMSD FNs in that their cores overlap but the rest of the molecule differs (*e.g.*, 3r17 and 2iiv, Fig. 5e). For the six SuCOS FNs, several have staggered rings when comparing the docked pose to the crystal pose (1k4g, 1o3p and 1kyn, Fig. 5f). For each example, there is a good contribution of shape overlap but not feature overlap to SuCOS (1k4g: shape = 0.68, features = 0.36; 1o3p: shape = 0.69, features = 0.35; 1kyn: shape = 0.70, features = 0.38). For shape overlap, the type of atom that overlaps is not considered, so for staggered heteroatomic rings, a relatively high shape overlap can be maintained. However, for shape plus feature overlap, there also needs to be exact matches of each feature type, so for staggered heterocycles the overlap of features can be poor. This explains why these staggered ring poses have relatively poor SuCOS scores.

Part III. Comparison of All-RMSD/MCS-RMSD, TvPLIF and SuCOS on the cross-docked larger ligand into the smaller ligand’s protein structure of the Malhotra and Karanicolas ligand pair set.

We refer to the scenario where a ligand is docked into the same protein but with a different conformation as “cross-docking”. Here, we docked the larger ligand, L_L , into its paired smaller ligand’s protein crystal structure, P_S^X (Fig. 1c). This simulates a virtual screening effort investigating potential fragment-hit follow ups by docking them into the fragment-hit’s protein structure. Each docking produced a number of docked poses, $PsL_L^{D,i}$, where i is the i^{th} docked pose produced for that cross-docking:

$$PsL_L^{D,i} = Dock(L_L, P_S^X) \quad (9)$$

Of the total 284 larger ligands, 242 were successfully docked into the smaller ligand’s protein structure, due to various errors as discussed in Part II such as the protein crystal structure having incomplete loops, or the ligand having unusual atom types. As described earlier, after generating

up to 10 conformers for each larger ligand, each conformer was docked using AutoDock Vina giving a total of 11,879 poses for the whole set, with an average of ~49 poses for each ligand. The MCS-RMSD was calculated for each pose of the larger, elaborated ligand comparing it to its corresponding smaller ligand, L_S^X , while All-RMSD was computed when comparing to itself, TvPLIF and SuCOS was also calculated for each docking pose with respect to both L_S^X and L_L^A .

The distributions of all metrics were all closer to L_S^X than to L_L^A (Supplementary Fig. S6). This is not unexpected, as comparing a larger ligand with its paired smaller ligand requires only part of the larger ligand to be similar. For example, for RMSD, only the maximum common substructure needs to overlap, while the rest of elaborated portion has no restrictions. Comparing the larger ligand with its crystal pose requires the whole structure to match.

We investigated whether picking poses using different metrics affected cross-docking success. For each of the 242 cross-dockings, one pose was kept for each metric. We compared the poses of the larger ligand with the crystal pose of the smaller ligand and retained the following: the one with the lowest MCS-RMSD, Eq. 10; the highest TvPLIF, Eq. 11; and the highest SuCOS, Eq. 12. The pose with the best AutoDock Vina affinity was also kept (Eq. 13). The results are summarized in Table 4.

$$Pose(RMSD_{best_to_smaller}) = Min(RMSD(L_L^{D,1}, L_S^X), RMSD(L_S^{D,2}, L_S^X), \dots RMSD(L_S^{D,i}, L_S^X)) \quad (10)$$

$$Pose(TvPLIF_{best_to_smaller}) = Max(TvPLIF(L_S^{D,1}, L_S^X), TvPLIF(L_S^{D,2}, L_S^X), \dots TvPLIF(L_S^{D,i}, L_S^X)) \quad (11)$$

$$Pose(SuCOS_{best_to_smaller}) = Max(SuCOS(L_S^{D,1}, L_S^X), SuCOS(L_S^{D,2}, L_S^X), \dots SuCOS(L_S^{D,i}, L_S^X)) \quad (12)$$

$$Pose(Aff_{best}^{Vina}) = Min(Aff^{Vina}(L_S^{D,1}), Aff^{Vina}(L_S^{D,2}), \dots Aff^{Vina}(L_S^{D,i})) \quad (13)$$

Table 4. Summary of cross-docking the larger ligand, L_L , from the MK dataset into the protein from the complex containing its paired smaller ligand, P_S^X .^a

Criterion	Number of poses	Number with All-RMSD < 2 Å to L_L^A (/242)	Number with TvPLIF > 0.6 to L_L^A (/242)	Number with SuCOS > 0.55 to L_L^A (/242)
Pose (RMSD _{best to L_S^X})	242	81	104	100
Pose (TvPLIF _{best to L_S^X})	242	60	116	82
Pose (SuCOS _{best to L_S^X})	242	76	109	106
Pose (Aff^{Vina} _{best})	242	62	81	78
MCS-RMSD < 2 Å to L_S^X	2,050	119	106	114
TvPLIF > 0.6 to L_S^X	2,338	111	121	113
SuCOS > 0.55 to L_S^X	1,949	114	111	117
Keeping all	11,879	140	153	134

^aInstead of using the native docking score, Aff^{Vina} , of the larger ligand from the docking, L_L^D to select a docked pose, it is possible to rank all of the docked poses of the larger ligand, L_L , by computing the MCS-RMSD, TvPLIF, and SuCOS against the smaller ligand's crystal structure, L_S^X . The success rates were computed against L_L^A . The criteria used to define a successful docking were All-RMSD < 2 Å, SuCOS > 0.55 and TvPLIF > 0.6. The success rates if all poses within a cutoff with respect to L_S^X are kept are also shown. The maximum success rate is also shown for each metric, if all poses are kept. Choosing one pose from each cross-docking leads to a much lower success rate across all metrics considered. A much higher recovery rate of the crystal structure of the larger ligand can be achieved if all poses within a given threshold are kept.

Using the definition of success as All-RMSD < 2 Å with respect to L_L^A , choosing the best pose by MCS-RMSD with respect to L_S^X achieved a success of 33% (81 cross-dockings) (Table 4). Similarly, choosing best by TvPLIF and SuCOS with respect to the L_S^X gives a success of 25% (60 cross-dockings) and 31% (76 cross-dockings) respectively. Interestingly, choosing best by Aff^{Vina} , gives a slightly better pose prediction than ranking by TvPLIF, with 26% success (62 cross-dockings). However, it should be noted that if all poses are kept, then 58% (140 cross-dockings) achieved at least one successful pose (Table 4).

Using this information, keeping all poses that satisfy a cutoff with respect to the smaller ligand pose may lead to greater pose prediction than keeping only the best pose. Hence, the following cutoffs were used to keep all poses that satisfy that cutoff: for MCS-RMSD, < 2 Å, for TvPLIF, > 0.6, for SuCOS, > 0.55. These criteria retained 17% (2,050), 20% (2,338) and 16%

(1,949) poses respectively. With these cutoffs, the success rates for MCS-RMSD, TvPLIF and SuCOS were 49% (119), 46% (111) and 47% (114) respectively.

This suggests that in a virtual screen where only the structural information of a smaller non-elaborated ligand is known, using only MCS-RMSD, TvPLIF or SuCOS to score against the smaller ligand to keep just one pose, may filter out poses which are actually closer to the larger ligand crystal pose. Alternatively, keeping all poses within a threshold of the smaller ligand crystal pose will give a better success of pose prediction.

Next, for each metric we considered the correlation of every docked pose with respect to the crystal structure of the smaller ligand, L_S^X , and the aligned crystal structure of the larger ligand, L_L^A (Fig. 6a). If a docking pose scores well with the smaller ligand crystal pose, then it should also score well with the larger ligand crystal pose and *vice versa*, provided the smaller ligand and its elaborated counterpart have a conserved binding mode — which was true in 86% of the cases studied by Malhotra and Karinicolos.¹ Therefore, the better the correlation, the better the metric should do at differentiating good poses.

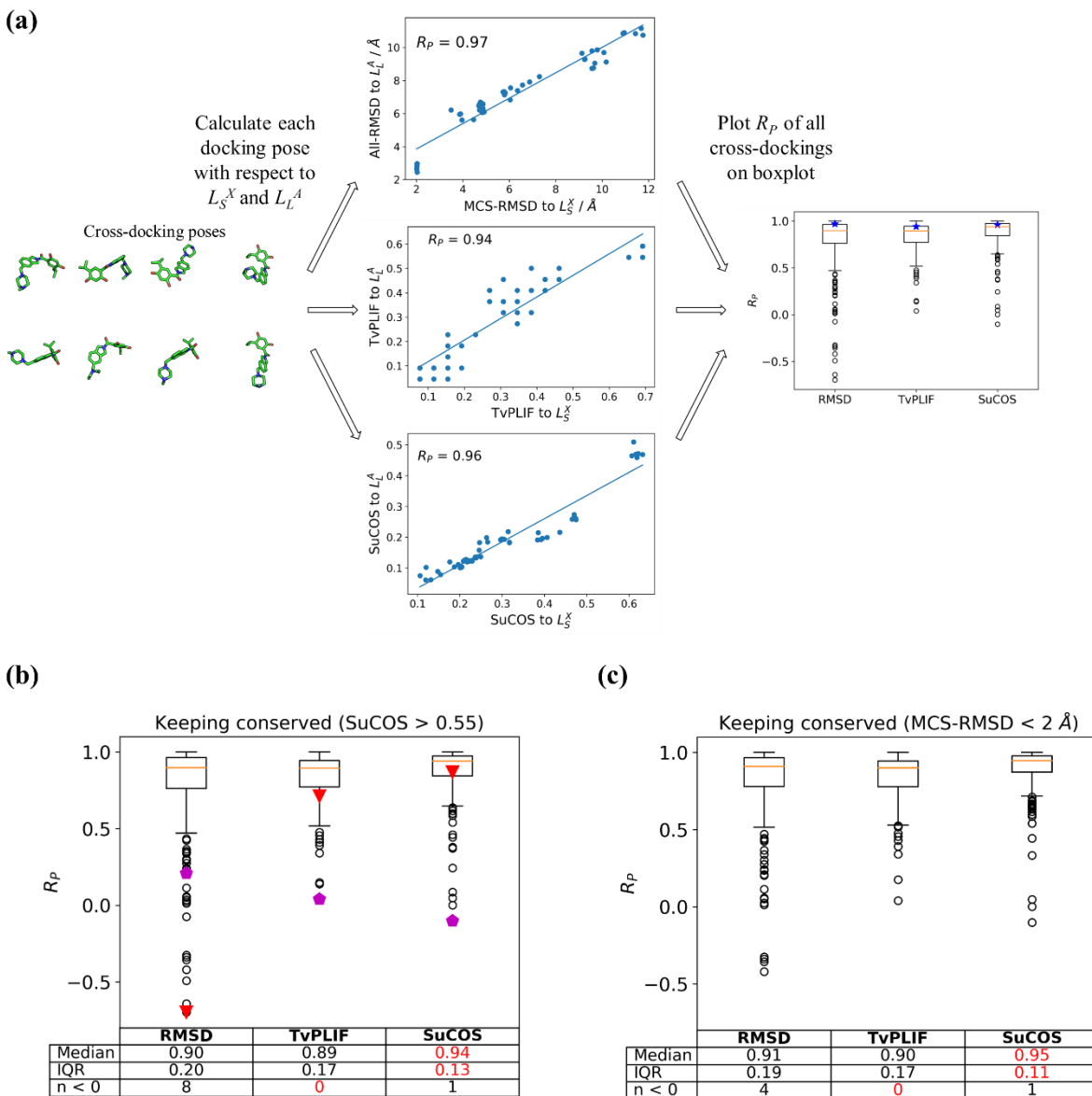


Fig. 6. SuCOS has the best correlations between docking poses of a larger ligand to its respective crystal pose and to its smaller counterpart crystal pose. (a) Schematic showing how boxplots (b) and (c) were created. For each cross-docking pose, the RMSD, TvPLIF, and SuCOS values were calculated with respect to the smaller and larger crystal ligand pose. These scores can be plotted on a scatter plot, with each point on the plots representing a single pose. The Pearson correlation coefficients, R_p , were then calculated for each metric for each cross-docking, comparing the

smaller ligand and the larger, elaborated ligand. The blue stars on the boxplots represent the Pearson correlation coefficients obtained for the example cross-docking shown. (b) Using SuCOS > 0.55 to define a conserved binding mode, the cross-docking pairs were filtered so only crystal ligands that showed a conservation of binding mode were kept. The collated Pearson correlation coefficients for each metric are shown on the boxplot. Four Pearson correlation coefficients were not included as there were fewer than nine outputted docking poses. The median, interquartile range (IQR) and number of negative Pearson correlation coefficients are shown in the table below the boxplot. The red triangles denote the RMSD outlier example shown in Fig. 7. The magenta pentagons denote the SuCOS outlier example shown in Supplementary S7. (c) Defining the conservation of binding mode with SuCOS does not bias the results. If the crystal ligand pairs are filtered according to MCS-RMSD (crystal ligand pairs with MCS-RMSD $< 2 \text{ \AA}$ kept), SuCOS still performs the best in terms of highest median and lowest IQR. The best values are highlighted in red.

Using SuCOS > 0.55 to define a conserved binding mode, the MK dataset was filtered to include only those elaborated ligands with a conserved binding mode. For each cross-docking, the All-RMSD to L_L^A was plotted against the MCS-RMSD to L_S^X for all the poses of that cross-docking. The distributions of Pearson correlation coefficients were plotted as boxplots and similar boxplots were drawn for TvPLIF and SuCOS (Fig. 6b).

A Pearson correlation coefficient of one indicates there is perfect correlation between the cross-docking poses and scoring to both the smaller ligand crystal pose and the larger ligand crystal pose. In most cases for all three metrics — RMSD, TvPLIF, and SuCOS — this correlation is near to one. In terms of median and interquartile range (IQR), however, SuCOS performed the best, with the highest median (0.94) and lowest IQR (0.13).

In order to avoid bias by using SuCOS to define the conservation of binding mode, we also performed the same analysis but filtered the MK dataset selecting only those poses with a conserved binding mode using the criterion MCS-RMSD $< 2 \text{ \AA}$ (Fig. 6c). SuCOS still performed better than RMSD and TvPLIF, in terms of both highest median and lowest IQR (0.95 and 0.11, respectively).

It is interesting to note there were a small number of cases with a negative Pearson correlation coefficient. SuCOS had one case with a negative R_p , whereas TvPLIF and RMSD had zero and eight respectively (Fig. 6b). In these cases, the cross-docking poses were visually inspected. For the one case of negative R_p for SuCOS (cross-docking of 3nus/3qd3), the corresponding R_p values for RMSD, TvPLIF and SuCOS are 0.21, 0.04 and -0.10 respectively (Supplementary S7). The docking achieved a pose within 2 \AA RMSD of the smaller and larger ligand. In this docked pose, the core rings are slightly staggered with respect to the crystal ligand pose. This resembles the SuCOS false negative examples shown in Fig. 5f. The ligand has multiple heteroatom rings, which may explain why these translations do not score well with SuCOS but better with RMSD. This cross-docking pose has good shape overlap but poor feature overlap with both the smaller and larger ligand. Consequently, any docking pose that corresponds to translation of rings will not score well.

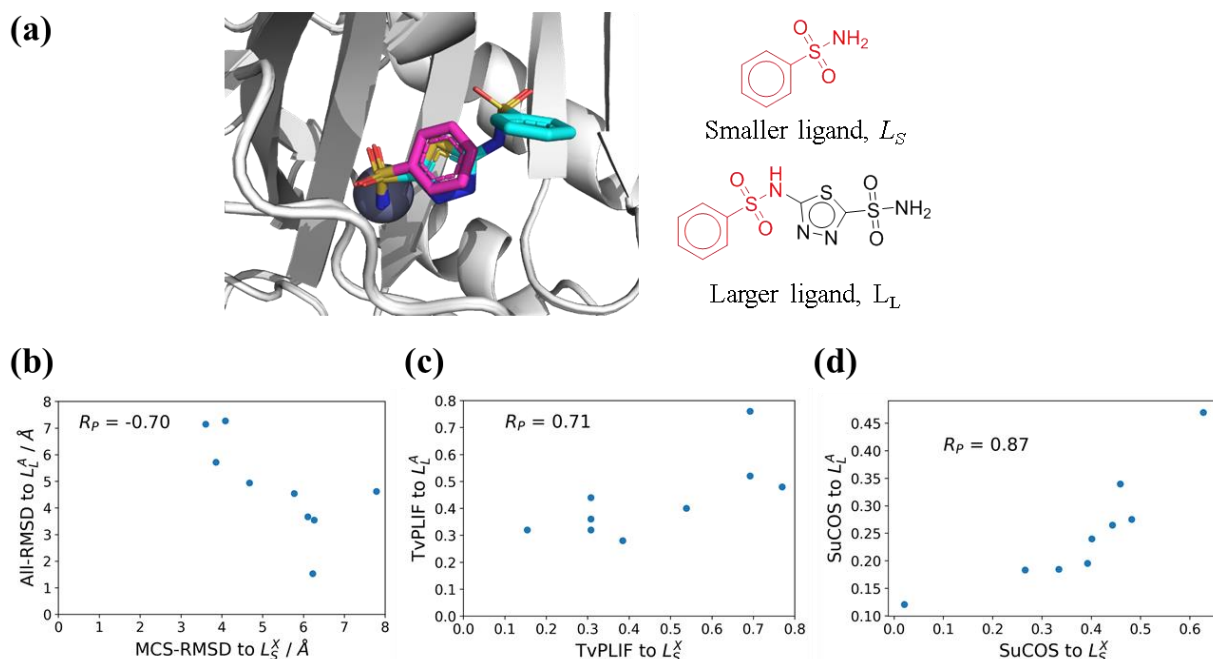


Fig. 7 Cross-docking of 2wej/3d8w: an example of a case where plotting the All-RMSD to the L_L^A against the MCS-RMSD to L_S^X for all the cross-docking poses gives a negative R_P . (a) 3D structure of smaller (pink) and larger (cyan) ligands in the binding sites of the aligned crystal structures of 2wej/3d8w. The 2D structure of the smaller and larger ligands are shown on the right with the MCS that was used to compute the RMSD highlighted in red. (b) Scatter plots showing the Pearson correlation coefficients of the cross-docking poses with the smaller and larger crystal structures for (b) RMSD, (c) TvPLIF, (d) and SuCOS.

The cross-docking of 2wej/3d8w is an example of where RMSD has a negative R_P (Fig. 7). This pair has good overlap of the crystal poses but the MCS-RMSD matches substructures that do not overlap (similar to 2xht/2vci in Table 2). This substructural mismatch resulted in the negative R_P found in the plot of RMSDs of the cross-docking poses to the large versus RMSD to the small (Fig. 7b).

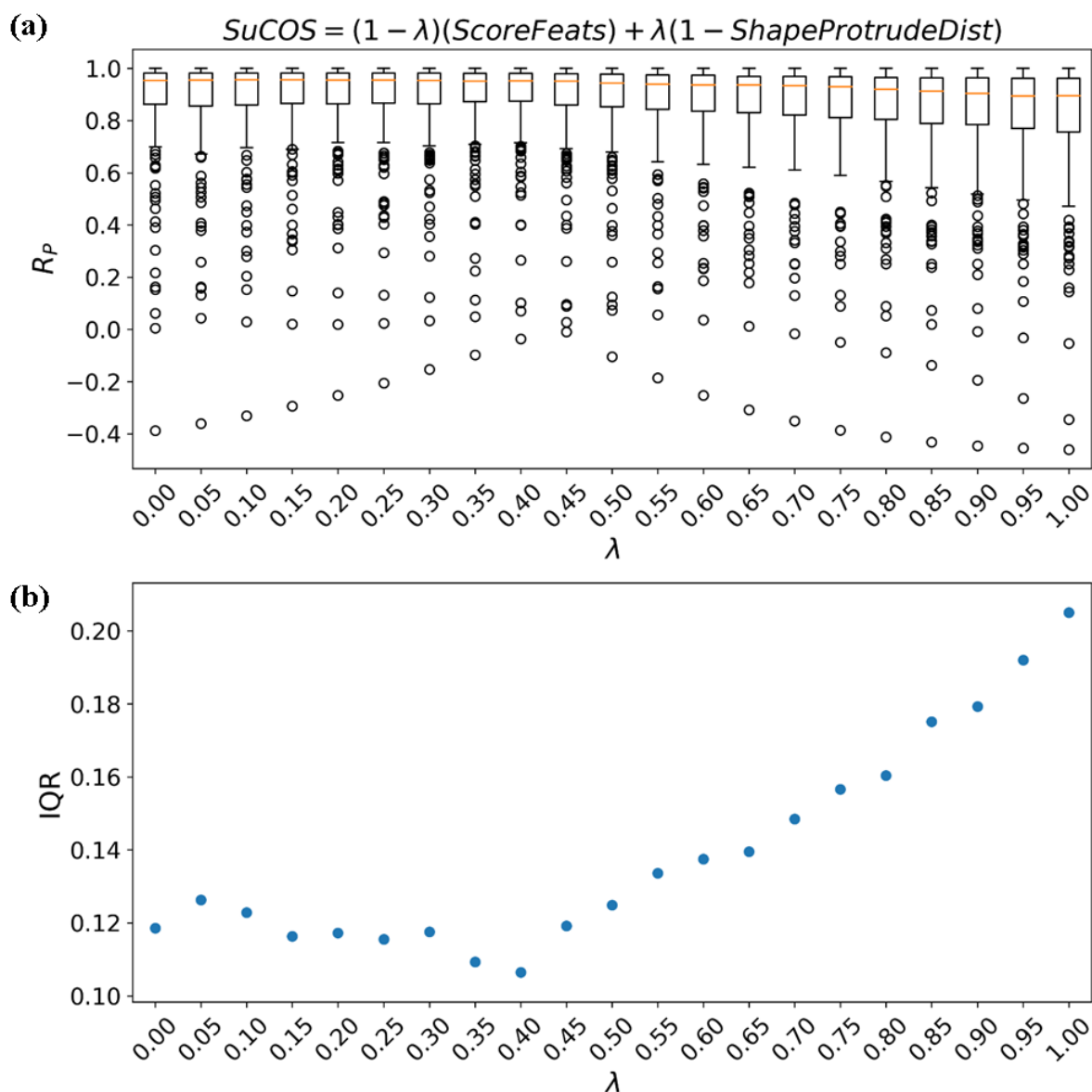


Fig. 8 Investigating the effect of altering the weights of feature overlap and shape overlap in SuCOS. (a) For each weighting, the distribution of Pearson correlation coefficients of the cross-dockings were plotted as boxplots. For the two extreme weightings, $\lambda = 0$ represents SuCOS using only chemical feature overlap and no shape overlap, while $\lambda = 1$ represents SuCOS with all shape overlap and no chemical feature overlap. (b) Plotting the IQR of the boxplots in (a) against the weights.

The weights of chemical feature overlap and shape overlap in the SuCOS metric have so far been assumed to be equal. We investigated the effect of altering the weights of each component of SuCOS. The weights of chemical feature overlap and shape overlap were changed from zero to one in increments of 0.05 and boxplots of the distributions of Pearson correlation coefficients of the cross-dockings were plotted for each (Fig. 8). The minimum IQR was found at a weight of $\lambda = 0.40$. This suggests that SuCOS may perform optimally with a slightly larger weighting of chemical feature overlap than shape overlap.

SuCOS performs better than AutoDock Vina using the DUD-E dataset

We compared the performance of SuCOS to that of AutoDock Vina for their ability to discriminate actives from decoys in the DUD-E dataset. The summary of the results is shown in **Fig. 9**. Scoring by SuCOS achieved a mean AUC across all targets of 0.775, whereas scoring by predicted Vina affinity achieved a corresponding value of 0.717 (Table 5 and Table S2 for all targets). In 64/102 cases SuCOS performed better than AutoDock Vina. SuCOS performed particularly well in some cases, with AUCs greater than 0.85, for 27 targets (FA7, TGFR1, KITH, PNPH, ADA, CAH2, MMP13, MAPK2, NRAM, CXCR4, PARP1, HMDH, ADA17, COMT, MET, FPPS, LKHA4, TYSY, HIVPR, GRIK1, DEF, SAHH, XIAP, WEE1, UROK, PUR2, THB). For example, SuCOS achieved an AUC of 0.882 for HIV-1 protease (**Fig. 9b**). SuCOS also performed better than AutoDock Vina for early enrichment, having a mean ROC enrichment factor at 1% across all targets, $RE_{1\%}$, of 27.254, compared to 10.797 for AutoDock Vina (Table 5 and Table S3-S6 for $RE_{0.5\%}$, $RE_{1\%}$, $RE_{2\%}$, $RE_{5\%}$, for all targets).

Table 5. Comparison of Mean AUC ROC and ROC Enrichment Values Across All DUD-E Targets when Scoring by AutoDock Vina and SuCOS.

Metric	AutoDock Vina	SuCOS
AUC ROC	0.717	0.775
RE _{0.5%}	15.210	48.161
RE _{1%}	10.797	27.254
RE _{2%}	7.599	15.866
RE _{5%}	5.016	8.061

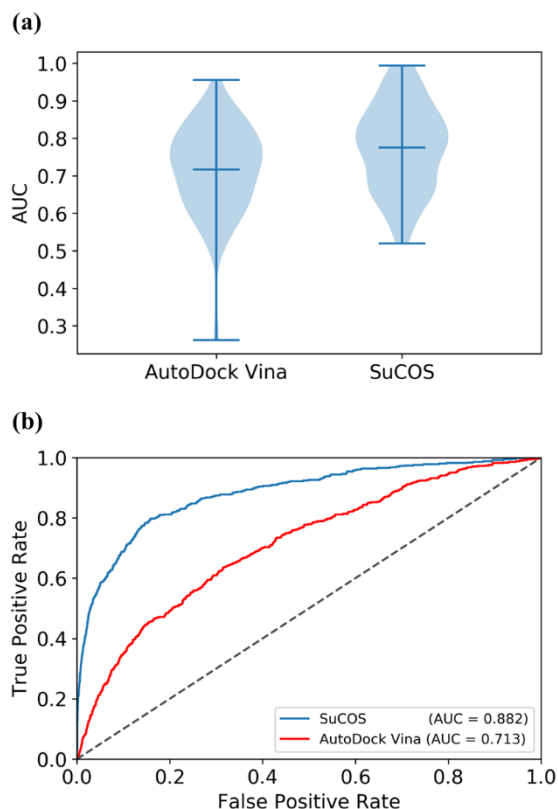


Fig. 9 (a) The distribution of area under the ROC curves for the 102 targets of the DUD-E dataset. The two violin plots show the different distributions obtained when scoring by scoring by the AutoDock Vina score versus SuCOS. For each plot, the maximum, minimum and mean are shown by the horizontal lines. (b) An example ROC plot: scoring by SuCOS achieves a better AUC (0.882) than scoring with the AutoDock Vina score for HIV-1 protease (0.713). The dotted line represents the performance of a random classifier.

CONCLUSIONS

We have compared three widely used metrics to investigate the conservation of binding mode of closely related smaller and larger ligands that bind to the same protein, namely: positional root

mean square deviation (RMSD); protein-ligand interaction fingerprint (PLIF) similarity; and shape-chemical feature overlap. For the shape-chemical feature overlap metric, we have introduced, SuCOS, an open-source RDKit-based metric. By investigating fragment-elaboration using the Malhotra-Karanicolas set, we have shown individual cases where each metric fails. RMSD is inappropriate to use when either molecule is pseudosymmetric, if multiple substructure matches are present, or there are bioisosteres. RMSD values depend heavily on the size of the molecules being compared, has no upper limit, and it is difficult to define a universal threshold for defining similarity. When comparing different molecules, multiple common substructure mismatches can sometimes occur, again invalidating the RMSD comparison.

PLIF similarity is heavily dependent on both the conformation of the protein and the ligand. Furthermore, is possible for a ligand to have a good PLIF similarity score but have a very different pose than the comparator. This means the pose of the ligand needs to be visually inspected, making it more time consuming and less straightforward to interpret than ligand-centric metrics. In addition, there is no universally accepted definition of protein-ligand interactions, which can affect the results greatly and what PLIF similarity threshold to use. Also, equal importance is given to each interaction type. On the other hand, for different conformations of a given protein, PLIF similarity can capture which interactions are conserved, whereas the other two metrics cannot do so explicitly.

Our focus here is on the comparison of poses of elaborated molecules and their non-elaborated counterparts, with the intention of using these results for structure-based virtual screening of elaborated molecules after a fragment soaking campaign. Any of the three metrics can be used to help choose a pose of an elaborated molecule given the crystal pose of its non-elaborated counterpart. However, much greater success could be obtained if we could accurately score all

the poses within a certain cutoff of the non-elaborated pose, *e.g.* RMSD < 2 Å to crystallographic binding mode of the fragment. The results from our studies on the MK dataset and the DUD-E benchmark suggest that docking algorithms can be improved by biasing the search using shape and chemical feature overlap to crystallographically-known ligands.

We have shown out of the three metrics, SuCOS obtains the best Pearson correlation coefficients when comparing poses of an elaborated molecule against its non-elaborated counterpart crystal structure and its true crystal pose. In a small number of cases with heterocyclic multi-ring systems, staggered conformations could result in poor SuCOS scores, but this could be obviated by adjusting the weights. We have shown that SuCOS is useful as both a conservation of binding mode metric, and as a tool for structure-based virtual screening. It is implemented using the open-source cheminformatics API, RDKit, hence making it accessible and easy to build upon. It is available on <https://github.com/susanhleung/SuCOS>.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI...

<SUMMARIZE SI>

AUTHOR INFORMATION

Corresponding Author

*Phone: +44 1865 281770. E-mail: garrett.morris@stats.ox.ac.uk.

ORCID

Susan H. Leung: 0000-0003-0917-0332

Frank von Delft: 0000-0003-0378-0017

Paul E. Brennan: 0000-0002-8950-7646

Garrett M. Morris: 0000-0003-1731-8405

ACKNOWLEDGMENTS

The authors would like to thank David Ryan Koes and Fergus Imrie for providing docked protein-ligand poses for the DUD-E dataset. This work was supported by funding from the Engineering and Physical Sciences Research Council (EPSRC) and the Medical Research Council (MRC) [grant number EP/L016044/1].

ABBREVIATIONS

COS, combined overlap score; FN, false negative; FP, false negative; MCS, maximum common substructure; PLIF, protein ligand interaction fingerprint; RMSD, root mean square deviation; R_p , Pearson correlation coefficient; TN, True Negative; TnPLIF, Tanimoto PLIF; TP, True Positive; TvPLIF, Tversky PLIF.

REFERENCES

- (1) Malhotra, S.; Karanicolas, J. When Does Chemical Elaboration Induce a Ligand to Change Its Binding Mode? *J. Med. Chem.* **2016**, acs.jmedchem.6b00725. <https://doi.org/10.1021/acs.jmedchem.6b00725>.
- (2) Drwal, M. N.; Jacquemard, C.; Perez, C.; Desaphy, J.; Kellenberger, E. Do Fragments and Crystallization Additives Bind Similarly to Drug-like Ligands? *J. Chem. Inf. Model.* **2017**, acs.jcim.6b00769. <https://doi.org/10.1021/acs.jcim.6b00769>.
- (3) Marcou, G.; Rognan, D. Optimizing Fragment and Scaffold Docking by Use of Molecular Interaction Fingerprints. *J. Chem. Inf. Model.* **2007**, 47 (1), 195–207. <https://doi.org/10.1021/ci600342e>.

- (4) Verdonk, M. L.; Ludlow, R. F.; Giangreco, I.; Rathi, P. C. Protein-Ligand Informatics Force Field (PLIFF): Toward a Fully Knowledge Driven “Force Field” for Biomolecular Interactions. *J. Med. Chem.* **2016**, *59* (14), 6891–6902. <https://doi.org/10.1021/acs.jmedchem.6b00716>.
- (5) Anighoro, A.; Bajorath, J. Three-Dimensional Similarity in Molecular Docking: Prioritizing Ligand Poses on the Basis of Experimental Binding Modes. *J. Chem. Inf. Model.* **2016**, *56* (3), 580–587. <https://doi.org/10.1021/acs.jcim.5b00745>.
- (6) Desaphy, J.; Raimbaud, E.; Ducrot, P.; Rognan, D. Encoding Protein-Ligand Interaction Patterns in Fingerprints and Graphs. *J. Chem. Inf. Model.* **2013**, *53* (3), 623–637. <https://doi.org/10.1021/ci300566n>.
- (7) Fu, D. Y.; Meiler, J. Predictive Power of Different Types of Experimental Restraints in Small Molecule Docking: A Review. *J. Chem. Inf. Model.* **2018**, *58* (2), 225–233. <https://doi.org/10.1021/acs.jcim.7b00418>.
- (8) Kumar, A.; Zhang, K. Y. J. Application of Shape Similarity in Pose Selection and Virtual Screening in CSARdock2014 Exercise. *J. Chem. Inf. Model.* **2016**, *56* (6), 965–973. <https://doi.org/10.1021/acs.jcim.5b00279>.
- (9) Onodera, K.; Satou, K.; Hirota, H. Evaluations of Molecular Docking Programs for Virtual Screening. *J. Chem. Inf. Model.* **2007**, *47* (4), 1609–1618. <https://doi.org/10.1016/j.ympev.2007.11.036>.
- (10) Plewczynski, D.; Łażniewski, M.; Augustyniak, R.; Ginalski, K. Can We Trust Docking Results? Evaluation of Seven Commonly Used Programs on PDBbind Database. *J. Comput.*

- Chem.* **2011**, 32 (4), 742–755. <https://doi.org/10.1002/jcc.21643>.
- (11) Warren, G. L.; Andrews, C. W.; Capelli, A. M.; Clarke, B.; LaLonde, J.; Lambert, M. H.; Lindvall, M.; Nevins, N.; Semus, S. F.; Senger, S.; et al. A Critical Assessment of Docking Programs and Scoring Functions. *J. Med. Chem.* **2006**, 49 (20), 5912–5931. <https://doi.org/10.1021/jm050362n>.
- (12) Allen, W. J.; Rizzo, R. C. Implementation of the Hungarian Algorithm to Account for Ligand Symmetry and Similarity in Structure-Based Design. *J. Chem. Inf. Model.* **2014**, 54 (2), 518–529. <https://doi.org/10.1021/ci400534h>.
- (13) Drwal, M. N.; Bret, G.; Perez, C.; Jacquemard, C.; Desaphy, J.; Kellenberger, E. Structural Insights on Fragment Binding Mode Conservation. *J. Med. Chem.* **2018**. <https://doi.org/10.1021/acs.jmedchem.8b00256>.
- (14) Bian, Y.; Feng, Z.; Yang, P.; Xie, X.-Q. Integrated In Silico Fragment-Based Drug Design: Case Study with Allosteric Modulators on Metabotropic Glutamate Receptor 5. *AAPS J.* **2017**, 19 (4), 1235–1248. <https://doi.org/10.1208/s12248-017-0093-5>.
- (15) Zaliani, A.; Boda, K.; Seidel, T.; Herwig, A.; Schwab, C. H.; Gasteiger, J.; Claußen, H.; Lemmen, C.; Degen, J.; Pärn, J.; et al. Second-Generation de Novo Design: A View from a Medicinal Chemist Perspective. *J. Comput. Aided. Mol. Des.* **2009**, 23 (8), 593–602. <https://doi.org/10.1007/s10822-009-9291-2>.
- (16) Durrant, J. D.; Amaro, R. E.; McCammon, J. A. AutoGrow: A Novel Algorithm for Protein Inhibitor Design. *Chem. Biol. Drug Des.* **2009**, 73 (2), 168–178. <https://doi.org/10.1111/j.1747-0285.2008.00761.x>.

- (17) Murray, C. W.; Verdonk, M. L.; Rees, D. C. Experiences in Fragment-Based Drug Discovery. *Trends Pharmacol. Sci.* **2012**, *33* (5), 224–232. <https://doi.org/10.1016/j.tips.2012.02.006>.
- (18) Ichihara, O.; Shimada, Y.; Yoshidome, D. The Importance of Hydration Thermodynamics in Fragment-to-Lead Optimization. *ChemMedChem* **2014**, *9* (12), 2708–2717. <https://doi.org/10.1002/cmdc.201402207>.
- (19) Temml, V.; Voss, C. V.; Dirsch, V. M.; Schuster, D. Discovery of New Liver X Receptor Agonists by Pharmacophore Modeling and Shape-Based Virtual Screening. *J. Chem. Inf. Model.* **2014**, *54* (2), 367–371. <https://doi.org/10.1021/ci400682b>.
- (20) Mysinger, M. M.; Carchia, M.; Irwin, J. J.; Shoichet, B. K. Directory of Useful Decoys, Enhanced (DUD-E): Better Ligands and Decoys for Better Benchmarking. *J. Med. Chem.* **2012**, *55* (14), 6582–6594. <https://doi.org/10.1021/jm300687e>.
- (21) Martin, A. C. R. Mapping PDB Chains to UniProtKB Entries. *Bioinformatics* **2005**, *21* (23), 4297–4301. <https://doi.org/10.1093/bioinformatics/bti694>.
- (22) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28* (1), 235–242.
- (23) Schrödinger, LLC, N. Y. *The PyMOL Molecular Graphics System, Version 2.1.0*.
- (24) Dolinsky, T. J.; Nielsen, J. E.; McCammon, J. A.; Baker, N. A. PDB2PQR: An Automated Pipeline for the Setup of Poisson-Boltzmann Electrostatics Calculations. *Nucleic Acids Res.* **2004**, *32* (Web Server issue), W665–7. <https://doi.org/10.1093/nar/gkh381>.
- (25) Olsson, M. H. M.; Søndergaard, C. R.; Rostkowski, M.; Jensen, J. H. PROPKA3: Consistent

- Treatment of Internal and Surface Residues in Empirical pK_a Predictions. *J. Chem. Theory Comput.* **2011**, 7 (2), 525–537. <https://doi.org/10.1021/ct100578z>.
- (26) Søndergaard, C. R.; Olsson, M. H. M.; Rostkowski, M.; Jensen, J. H. Improved Treatment of Ligands and Coupling Effects in Empirical Calculation and Rationalization of pK_a Values. *J. Chem. Theory Comput.* **2011**, 7 (7), 2284–2295. <https://doi.org/10.1021/ct200133y>.
- (27) Morris, G. M.; Huey, R.; Lindstrom, W.; Sanner, M. F.; Belew, R. K.; Goodsell, D. S.; Olson, A. J. AutoDock4 and AutoDockTools4: Automated Docking with Selective Receptor Flexibility. *J. Comput. Chem.* **2009**, 30 (16), 2785–2791. <https://doi.org/10.1002/jcc.21256>.
- (28) MolVS: Molecule Validation and Standardization — MolVS 0.0.9 documentation <http://molvs.readthedocs.io/en/latest/> (accessed Jun 13, 2017).
- (29) O’Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R. Open Babel: An Open Chemical Toolbox. *J. Cheminform.* **2011**, 3 (10), 1–14. <https://doi.org/10.1186/1758-2946-3-33>.
- (30) Trott, O.; Olson, A. AutoDock Vina: Improving the Speed and Accuracy of Docking with a New Scoring Function, Efficient Optimization and Multithreading Oleg. *J. Comput. Chem.* **2010**, 31 (2), 455–461. <https://doi.org/10.1002/jcc.21334>.AutoDock.
- (31) RDKit, Version 2018.03.1. Open-source cheminformatics 2018.
- (32) Deng, Z.; Chuaqui, C.; Singh, J. Structural Interaction Fingerprint (SIFt): A Novel Method for Analyzing Three-Dimensional Protein–Ligand Binding Interactions. *J. Med. Chem.*

- 2004**, 47 (2), 337–344. <https://doi.org/10.1021/jm030331x>.
- (33) Radifar, M.; Yuniarti, N.; Istyastono, E. P. PyPLIF : Python-Based Protein-Ligand Interaction Fingerprinting Abstract : Background : Methodology : *Bioinformatics* **2013**, 9 (6), 325–328. <https://doi.org/10.6026/97320630009325>.
- (34) Jubb, H. C.; Higuieruelo, A. P.; Ochoa-montaña, B.; Pitt, W. R.; Ascher, D. B.; Blundell, T. L. Arpeggio : A Web Server for Calculating and Visualising Interatomic Interactions in Protein Structures. *J. Mol. Biol.* **2017**, 429 (3), 365–371. <https://doi.org/10.1016/j.jmb.2016.12.004>.
- (35) Ragoza, M.; Hochuli, J.; Idrobo, E.; Sunseri, J.; Koes, D. R. Protein–Ligand Scoring with Convolutional Neural Networks. *J. Chem. Inf. Model.* **2017**, 57 (4), 942–957. <https://doi.org/10.1021/acs.jcim.6b00740>.
- (36) Imrie, F.; Bradley, A. R.; van der Schaar, M.; Deane, C. M. Protein Family-Specific Models Using Deep Neural Networks and Transfer Learning Improve Virtual Screening and Highlight the Need for More Data. *J. Chem. Inf. Model.* **2018**, 58 (11), 2319–2330. <https://doi.org/10.1021/acs.jcim.8b00350>.
- (37) DUD-E docked poses http://bits.csb.pitt.edu/files/docked_dude.tar (accessed May 2, 2019).
- (38) Koes, D. R.; Baumgartner, M. P.; Camacho, C. J. Lessons Learned in Empirical Scoring with Smina from the CSAR 2011 Benchmarking Exercise. *J. Chem. Inf. Model.* **2013**, 53 (8), 1893–1904. <https://doi.org/10.1021/ci300604z>.
- (39) Nicholls, A. What Do We Know and When Do We Know It? *J. Comput. Aided. Mol. Des.* **2008**, 22 (3–4), 239–255. <https://doi.org/10.1007/s10822-008-9170-2>.

- (40) Jain, A. N.; Nicholls, A. Recommendations for Evaluation of Computational Methods. *J. Comput. Aided. Mol. Des.* **2008**, 22 (3–4), 133. <https://doi.org/10.1007/S10822-008-9196-5>.
- (41) ROCS, Version 3.2.0.3. OpenEye Scientific: Sante Fe, NM 2015.
- (42) Bursulaya, B. D.; Totrov, M.; Abagyan, R.; Brooks, C. L. Comparative Study of Several Algorithms for Flexible Ligand Docking. *J. Comput. Aided. Mol. Des.* **2003**, 17 (11), 755–763. <https://doi.org/10.1023/B:JCAM.0000017496.76572.6f>.
- (43) Marcou, G.; Rognan, D. Optimizing Fragment and Scaffold Docking by Use of Molecular Interaction Fingerprints. *J. Chem. Inf. Model.* **2007**, 47 (1), 195–207. <https://doi.org/10.1021/ci600342e>.
- (44) Anighoro, A.; Bajorath, J. Binding Mode Similarity Measures for Ranking of Docking Poses: A Case Study on the Adenosine A2A Receptor. *J. Comput. Aided. Mol. Des.* **2016**, 30 (6), 447–456. <https://doi.org/10.1007/s10822-016-9918-z>.
- (45) Liu, J.; Su, M.; Liu, Z.; Li, J.; Li, Y.; Wang, R. Enhance the Performance of Current Scoring Functions with the Aid of 3D Protein-Ligand Interaction Fingerprints. *BMC Bioinformatics* **2017**, 18 (1), 343. <https://doi.org/10.1186/s12859-017-1750-5>.
- (46) Ramírez, D.; Caballero, J. Is It Reliable to Take the Molecular Docking Top Scoring Position as the Best Solution without Considering Available Structural Data? *Molecules* **2018**, 23 (5), 1–17. <https://doi.org/10.3390/molecules23051038>.
- (47) Hawkins, P. C. D.; Nicholls, A. Conformer Generation with OMEGA: Learning from the Data Set and the Analysis of Failures. *J. Chem. Inf. Model.* **2012**, 52 (11), 2919–2936.

<https://doi.org/10.1021/ci300314k>.

- (48) Hawkins, P. C. D.; Warren, G. L.; Skillman, A. G.; Nicholls, A. How to Do an Evaluation: Pitfalls and Traps. *J. Comput. Aided. Mol. Des.* **2008**, *22* (3–4), 179–190.
<https://doi.org/10.1007/s10822-007-9166-3>.