

Cite this: DOI: 10.1039/xxxxxxxxxx

# Antibody-Antigen Complex Modelling in the Era of Immunoglobulin Repertoire Sequencing

Matthew I. J. Raybould,<sup>a‡</sup> Wing Ki Wong,<sup>a‡</sup> and Charlotte M. Deane<sup>a</sup>

Received Date

Accepted Date

DOI: 10.1039/xxxxxxxxxx

www.rsc.org/journalname

The natural immune repertoire can be a useful guide to antibody discovery against any given target. However, the large volume of immunoglobulin gene sequencing data necessitates the rational prioritisation of possible binders for experimental validation. Where other known binders exist, sequence similarity is used to infer binding, but this neglects alternative binding modes to the same epitope, and cannot identify antibodies that bind to different epitopes. In this review, we summarise the state-of-the-art of high-throughput antibody-antigen complex modelling. Given the millions of natural antibody sequences now available, this pipeline attempts to predict whether, and if so how, each antibody binds to a particular antigen's surface. We cover the current paradigm (antibody and antigen structural modelling, followed by binding site prediction, followed by molecular docking), discussing how existing algorithms can deal with this magnitude of data by balancing accuracy with computational efficiency, and identifying areas where further developments are required to improve performance.

## 1 Introduction

Since 2009, Next-Generation Sequencing of immunoglobulin gene repertoires (Ig-seq) has provided samples of natively-expressed antibody chains<sup>1</sup>, with the aim of better understanding the constitution of the immune system. Over the following decade, the number and magnitude of Ig-seq datasets has exponentially increased and they now cover a variety of organisms and immune states<sup>2</sup>. For example, the Observed Antibody Space (OAS) database<sup>2</sup> currently totals 60 studies, and contains over one billion annotated Ig-seq reads across 6 species (valid as of 30<sup>th</sup> March, 2019). The latest update included an Ig-seq study by Briney *et al.*<sup>3</sup>, which alone released over 300M productive human heavy chain antibody sequences into the public domain.

This influx of data is allowing us to analyse how naïve and mature antibody sequences differ, improving our understanding of the dynamics of the adaptive immune response. However, if we wish to move beyond this surface information to accurately describe how they differ structurally, we will need methods that can rapidly determine the binding sites represented within each dataset.

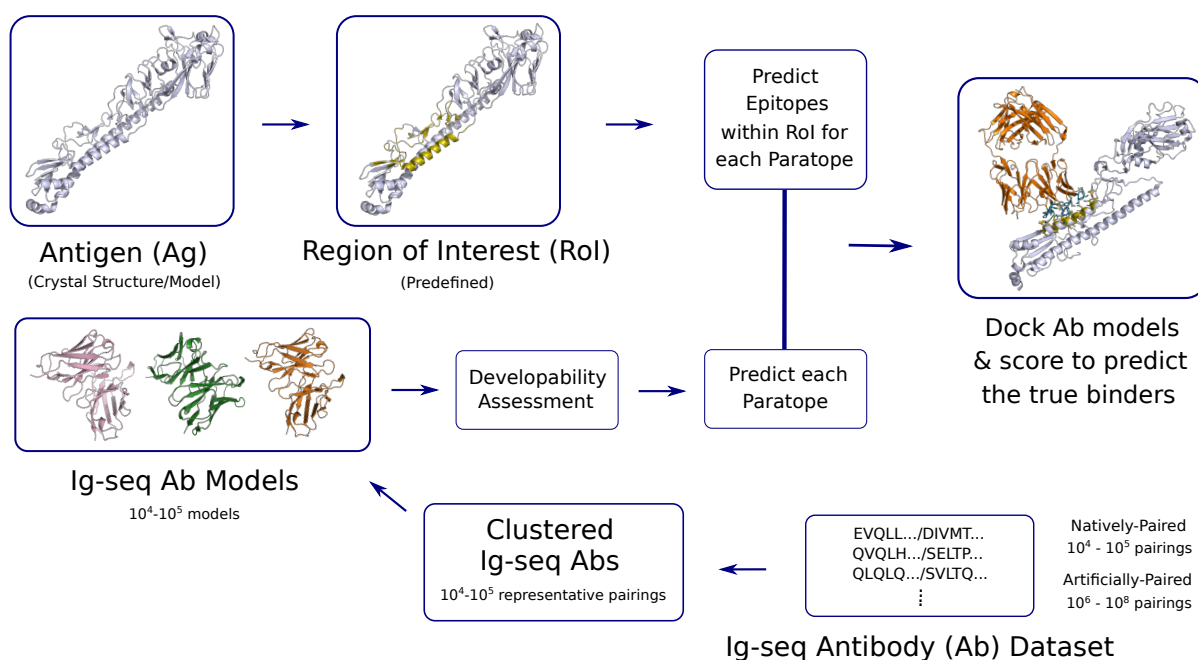
Additionally, natively-expressed human antibody sequences potentially represent good starting points in therapeutic development against any given target ('antigen'). Monoclonal antibodies (mAbs) derived from these variable domains, particularly if coupled in natural heavy-light chain pairings, ought to have reasonable expression and a lower risk of immunogenicity. However,

it is not feasible to perform experimental binding characterisation on every productive antibody that could be isolated from an Ig-seq experiment. To achieve a more tractably-sized subset for experimental validation, Ig-seq datasets could be probed for structures that are likely to both bind to a given target (increasingly targetting a particular surface patch, 'epitope'<sup>4</sup>), and have intrinsic characteristics compatible with therapeutic manufacture and storage<sup>5</sup>.

Co-crystal complexes are the 'gold standard' in binding site characterisation, but they are difficult and costly to obtain. As of March 2019, fewer than 2,500 antibody-antigen (Ab-Ag) complexes have been publicly released<sup>6</sup>. These structures can be used to suggest which Ig-seq antibodies are likely to bind in a similar manner by considering sequence similarity, assuming that antibodies with sequence-similar binding sites ('paratopes') will engage the same epitope on the antigen. However, this technique will always fail to recognise antibodies with considerably different paratopes that can bind to the same epitope by exploiting a different binding mode<sup>7</sup>, as well as those that bind to new, uncharacterised epitopes.

High-throughput Ab-Ag complex modelling offers a computational route to deducing the small subset of Ig-seq antibodies that can bind to a given antigen. By combining *in silico* procedures (currently a sequential pipeline of structural modelling, Ab-Ag interface prediction, and molecular docking; Fig. 1), antibodies complementary to each particular region of the antigen surface could be identified. Such methodology would not only be useful in computationally proposing antigen-binders directly from Ig-seq

‡ These authors contributed equally to this work.



**Fig. 1** A schematic showing the proposed paradigm for high-throughput antibody-antigen (Ab-Ag) complex modelling of Ig-seq data. An Ig-seq dataset of natively-paired or artificially-paired sequences is chosen. A subset of antibodies can be chosen from the Ig-seq repertoire, using methods such as sequence clustering to obtain a more computationally tractable number. Each antibody structure is then predicted using a rapid homology modelling tool. At this point, it is sensible to run *in silico* developability assessment tools to highlight which antibodies may be particularly prone to developability issues. Separately, a structure of the antigen of interest is collected or modelled, and a key binding region of interest is determined. The most probable paratope (antibody binding component) and epitope (antigen binding component, within the region of interest) are predicted for each antibody. Finally, docking and scoring are used to determine which antibodies are likely to be complementary, and to suggest how complementary antibodies might bind to the target. This pipeline can be repeated for multiple regions of interest on the antigen.

data, but also in the scenario where binding antibody sequences are known through *in vitro* analysis, but only those that engage a particular epitope, or those that target the broadest range of epitopes, are desired for further development.

In this review, we will discuss the tools available to perform each step of Ab-Ag complex modelling. We will suggest how they may be used in concert to generate as reliable a prediction as possible, while retaining optimal computational efficiency to ensure that the protocol is compatible with large Ig-seq datasets. We will also consider which steps are most likely to introduce inaccuracies, and detail either how algorithmic readjustment may be able to resolve the issue, or describe where further development is essential to achieve more reliable results.

## 2 Immunoglobulin repertoire sequencing

An antibody variable domain contains an interacting heavy and light chain, whose internal structures can each be subdivided into four framework regions (FWRs) connecting three loops (Complementary Determining Regions, CDRs), the boundaries between which can be defined differently<sup>8–11</sup>. To a reasonable approximation, antibodies engage antigens using a combination of their CDR loops, often dominated by the hypervariable CDRH3 loop<sup>12</sup>. It follows that knowledge of the variable domain composition alone should be sufficient to predict any antibody's specificity<sup>13</sup>. Accordingly, Ig-seq experiments tend to focus on accurately capturing the sequence of the CDRH3 loop, the heavy variable do-

main ( $V_H$ ), or both heavy and light ( $V_L$ ) variable domains<sup>2</sup>. Precisely which chain/region is sequenced depends on the experimental protocol used.

While light chains are necessary stabilising components of antibodies, they are much less diverse in sequence, and are considered to contribute less often to binding affinity and specificity. Ig-seq studies therefore less frequently include light chain primers, leading to an imbalance in sequencing data (there are 651M human heavy chain reads against 53.6M human light chain reads in the OAS database<sup>2</sup> as of 30<sup>th</sup> March, 2019). Studies with pooled heavy and light chain primers are often performed on pooled B cells, which results in a loss of the native pairing. If complete representations of these binding sites are required, a computational protocol must be employed to hypothesise feasible pairings<sup>5</sup>. Increasingly, single-cell sequencing is providing immune repertoire snapshots where the native heavy and light chain pairings are preserved<sup>14,15</sup>. Currently limited to samples on the order of  $10^4$ - $10^5$  antibodies, this technique is likely to scale up rapidly over the coming years, and will hopefully soon provide sufficiently comprehensive immune repertoire snapshots for direct use in antibody drug discovery.

Nanobodies (Nbs), which can penetrate deeper into tumours than antibodies and target intracellular antigens, are also of increasing therapeutic interest<sup>16</sup>. These immune proteins, derived from camelids and cartigenous fish, have the advantage of only comprising a single chain (a  $V_{HH}$  domain), circumventing the

need to derive productive pairings or perform single-cell sequencing. In principle, a similar, though bespoke, approach to Ab-Ag complex modelling could be proposed to predict Nb-Ag complex interactions. However, although an initial study on the binding properties of nanobodies has recently been released<sup>17</sup>, we have far less sequence and structural data for nanobodies than we do for antibodies: only one Nb Ig-seq study<sup>18</sup> is recorded in the OAS database, containing 1.15M reads, and only 394 Nb structures (297 of which are complexes) are recorded in SAbDab<sup>6</sup> as of 30<sup>th</sup> March, 2019. The remainder of this review will therefore focus on the more immediately achievable goal of Ab-Ag complex modelling.

The number of productive immunoglobulins proposed by even a single Ig-seq experiment is prohibitively large to run every antibody sequence through binding analysis. Instead, these large Ig-seq datasets must be filtered to balance the competing factors of diversity and experimental or computational tractability. This could first be achieved by eliminating sequences without a satisfactory structural template for every CDR loop (see Section 3), followed by sequence identity/similarity clustering to identify representative surviving chains. After productive V<sub>H</sub>-V<sub>L</sub> pairings are proposed, additional low-resolution structural clustering could be performed, for example by comparing the distances between proposed CDR loop and/or V<sub>H</sub>-V<sub>L</sub> orientation templates<sup>5</sup>. Longitudinal vaccination Ig-seq studies could also be harnessed to identify the subset of sequences or clusters enriched post-immunisation<sup>19,20</sup>.

### 3 Antibody modelling

Experimentally determining crystal structures of antibodies is currently costly and low-throughput. Since the first variable domain structure was deposited in the PDB<sup>21</sup> in 1976, fewer than 3,400 variable region structures have been solved, many of which are entirely redundant in sequence<sup>6</sup>. If we are to make use of the quantity of sequence data now available from Ig-seq analyses, it is necessary to employ tools that can computationally derive features of each variable domain structure from its sequence. This type of structural characterisation has already been shown to be a powerful way to analyse Ig-seq datasets; Kovaltsuk *et al.*<sup>22</sup> explain how it more reliably distinguishes similar and dissimilar binding sites.

In the paper by Krawczyk *et al.*<sup>23</sup>, they demonstrate that, to a first approximation, it is possible to capture the geometric properties of Ig-seq antibody binding sites without constructing a full model. For example, the SAAB protocol<sup>23</sup> harnessed the fact that the CDRH1 and CDRH2 loops (and CDRL1, CDRL2 and CDRL3 loops, if applied on light chain data) typically adopt a limited ensemble of backbone conformations ('canonical forms'<sup>9</sup>), and that these forms can be rapidly and reliably predicted from sequence<sup>11,24</sup>. Owing to the genetic mechanisms of combinatorial and junctional diversity, in addition to somatic hypermutation, CDRH3 varies far more in sequence and in length than the other CDRs, and so its structure has to be predicted by other methods. SAAB, for example, proposes CDRH3 structures if the target sequence can be matched to another CDRH3 of known structure<sup>25,26</sup>. Coupling these techniques together allows for an effi-

cient description of binding site geometry for a large portion of the dataset. This approximate representation was used to identify potential sequence dissimilar binders to a known influenza epitope<sup>23</sup>, and a similar approach was utilised to cluster Ig-seq data en route to creating a set of models with diverse binding sites<sup>5</sup>.

To reliably model an antibody-antigen interface, models of the entire antibody variable domain are required to atomistic detail. There are many software packages capable of performing this task<sup>27–36</sup>. Some tools are freely available to all users<sup>27–30</sup>, others are free to use under an academic license, but a paid subscription applies for commercial users<sup>31</sup>, and still others require the purchase of either an academic or commercial license<sup>32–36</sup>.

Most of these software packages reliably achieve close to sub-ångström accuracy for the more sequence homogenous regions among antibodies. These include the four framework regions (FWR1-4), and, despite somatic hypermutation, the canonical CDR loops. Such accuracy is achieved using homology modelling approaches<sup>25,26,37–42</sup>. These methods use a template region of known structure and similar sequence to infer the structure of the target region. Each tool has its own curated database of framework, CDR, and orientation templates, all of which harness known antibody structures from the PDB<sup>21</sup>.

Framework region templates are typically selected based solely on maximal sequence identity or similarity to the target. Either a single template antibody can be used for both chains, a separate template can be chosen for each chain, or templates can be chosen separately for each intra-chain framework region. Attention is then given to predicting a relative V<sub>H</sub> and V<sub>L</sub> domain orientation<sup>43–45</sup>. There are many ways to define the interface topology, for example, Dunbar *et al.*<sup>44</sup> in their ABangle software define a V<sub>H</sub>-V<sub>L</sub> orientation by a set of five angles and one distance parameter between planes through each variable domain. If all templates come from the same parent antibody, then the parent's V<sub>H</sub>-V<sub>L</sub> orientation is usually imparted directly to the model structure. However, if the templates come from an array of parent antibodies, then interface parameters must be assigned algorithmically. One computationally inexpensive solution is to predict which residues are likely to be involved in the V<sub>H</sub>-V<sub>L</sub> interface, and evaluate the sequence identity/similarity across these positions to known structures, with the closest match bestowing its interface parameters to the model<sup>27</sup>. In contrast to the FWRs, the CDRs are always considered separately, and any combination of loop length, canonical form, sequence similarity, and anchor residue distance is used to select the best template.

Side chain conformations can be considered either as each region is modelled, or after the entire backbone is constructed. Opinions differ as to whether retaining the side chain conformations of residues present in both target and template, or remodelling all side chains in the context of the new model gives the best performance. Side chains can be modelled using either proprietary or freely-available<sup>46–50</sup> tools, all of which combine an energy function with a rotamer library. Most tools were developed for general protein side chain modelling<sup>46–49</sup>, while the PEARS tool is designed specifically to model antibody side chains<sup>50</sup>.

The largest difference between modelling tools tends to be in

their treatment of CDRH3, whose structure tends to be predicted poorly<sup>12</sup> as its conformational search space is much greater than any other CDR. Whenever a close template is available, homology modelling allows CDRH3 structure prediction in a far shorter time-frame than *ab initio* methods (seconds, rather than hours) and can yield sub-ångström accuracy<sup>51</sup>. However, as our coverage of CDRH3 structural space is so limited, it is often the case that no suitable templates can be found, or that models based on a single template are of poor quality. To improve on this, varying degrees of *ab initio* modelling can be performed.

*Ab initio* modelling<sup>52–57</sup> requires no template, instead generating decoys from scratch and energy minimising them. Initial guesses of model coordinates are usually generated by some stochastic algorithm, for example by randomly sampling the  $\phi$  and  $\psi$  angles of each residue's Ramachandran space. These coordinates are then perturbed to minimise some energy function. In particular, *ab initio* loop modelling can either be performed by growing from one end of the loop, with geometric constraints to ensure loop closure, or by guessing an initial closed loop backbone conformation and then minimising the energy.

Some *ab initio* algorithms, such as Rosetta's KIC<sup>52</sup>, will then re-evaluate  $V_H$ - $V_L$  orientation, and repeat decoy generation if the interface changes significantly. Such refinement can improve prediction accuracy, but takes hundreds of CPU hours<sup>51</sup>. Algorithms that combine *ab initio* predictions and loop fragment templates, either in a consensus fashion<sup>25</sup> or as a hybrid method<sup>51,58–60</sup>, can dramatically reduce prediction times. All proposed decoys are ranked according to a scoring function, and the highest ranked, or a consensus of multiple top-rankers, is inserted into the model. CDRH3 modelling methods are reviewed in greater detail by Marks and Deane<sup>61</sup>.

Once a model has been generated, some tools, such as ABody-Builder<sup>27</sup>, provide an estimate of regional model accuracy. These estimates are helpful as they can either be harnessed downstream to select appropriate docking parameters (see Section 6), or used to filter out low-confidence models (see Section 4).

Modelling Ig-seq data necessitates a protocol that can balance speed with accuracy. Modelling tools that attempt to homology model every loop are by far the most high-throughput methods, taking fractions of CPU minutes per antibody, if all loops have good templates. In contrast, algorithms that will only model CDRH3 loops in an *ab initio* fashion need to generate many decoys, and prediction can take multiple CPU hours per antibody.

This naturally leads to two possibilities: (a) only model Ig-seq antibodies for which every CDR has an adequate template, discarding the others, or (b) model Ig-seq antibody loops by homology as a default, and use *ab initio*/hybrid approaches only if a particular loop sequence (usually, but not always CDRH3) is completely devoid of templates. Solution (a) is likely to be rapid yet result in an under-sampling of antibodies with long CDRH3 loops, while solution (b) will have better coverage but take much longer to compute and may be prohibitive for large human Ig-seq datasets. While undersampling longer CDRH3s may not be too problematic for most therapeutic targets, given the length biases currently observed across approved mAb therapies<sup>5</sup>, antibodies against HIV, for example, require long CDRH3 loops for comple-

mentarity<sup>62</sup>. It is therefore important to improve our ability to homology model long CDRH3 loops, and a targeted effort to obtain more crystal structures of antibodies containing such loops would be highly beneficial to the field.

At this stage in the pipeline, the decision could be made to *in silico* screen the antibody models for their intrinsic propensity to possess 'developability issues'<sup>5,63–65</sup>, such as aggregation, immunogenicity, or polyspecificity. The highest-risk sequences could then either be removed, or retained alongside a flag to highlight them as likely to need developability-related refinement.

## 4 Paratope prediction

The next step of Ab-Ag complex modelling involves identifying which residues are likely to comprise the antibody binding site (paratope). Accurate paratope definitions are necessary for relevant high-speed docking results. As previously discussed, antigen binding typically involves residues in the CDR loops. On average, across all definitions, CDRs capture 80% of the antigen-binding residues, as some regions of the framework (in particular FWR3) tend to lie within binding distance, and the CDRs also contain many residues not involved in binding<sup>66,67</sup>. Several computational methods exist to accurately predict paratopes with sufficient rapidity for use on Ig-seq datasets.

The majority of these methods take only a variable domain sequence as input. Kunik *et al.*<sup>67</sup> developed their Paratome software by harnessing sequence and structural data to estimate the energetic importance of each structurally-conserved antibody position to antigen binding. By incorporating binding residue patterns into a random forest model, proABC can also predict paratope residues on any given input antibody sequence<sup>68</sup>. Most recently Liberis *et al.*<sup>69</sup> have built Parapred, which trains a neural network on non-redundant Ab-Ag complexes to predict paratope residues, achieving an impressive ROC AUC of  $0.878 \pm 0.004$  across 10-fold cross validation. They also show that this improvement in prediction accuracy translated into better subsequent docking performance, strongly suggesting that further improvements in paratope prediction will return tangible benefits to the reliability of Ab-Ag complex modelling.

The Antibody i-Patch software<sup>70</sup> utilises structures of both the antibody and antigen to generate its paratope prediction. It assigns a binding likelihood score to each input antibody residue based on the frequency of triplets of binding residues observed across antibody-antigen crystal structure interfaces. As it takes into account the structure of both partners, this tool should return more bespoke results for the cognate antigen of interest. In the context of Ig-seq modelling, where we must use antibody (and potentially antigen) models, it is crucial that these models are of high enough quality to add value rather than noise to the prediction. Antibody modellers that give predictions of model accuracy<sup>27</sup> are particularly useful here, as the choice can be made to either remove these models from the dataset, or refine them with more sophisticated and computationally-expensive methods.

## 5 Antigen modelling and epitope prediction

Antibodies can bind to virtually any class of antigen: nucleic acids, haptens, peptides or proteins. Ideally, a crystal structure of

the antigen of interest is available. However, if the structure must be predicted, protein antigens pose a unique challenge owing to their potential to be extremely large and diverse. Macromolecular structure prediction is a highly active field, and for some antigens it may not yet be possible to propose a sufficiently accurate model. The Critical Assessment of Structure Prediction (CASP) is a biennial competition in which participants are asked to predict a range of novel protein structures from sequence, to ascertain the progress made by the field. Reports from the latest iteration (<http://predictioncenter.org/casp13/index.cgi>) show that improvements continue to be made across all categories of protein modelling. Well-characterised protein sub-classes remain much easier to model accurately, primarily as they benefit from an abundance of template fragments. For other targets, where template-free approaches must be used, accuracy is considerably lower, though has recently improved through our ability to better predict residue-residue contacts.

An epitope of an antigen is defined as a subset of its surface to which an antibody can bind. Epitopes fall into two categories: linear and conformational.

Linear epitopes are contiguous polypeptide chains, and are relatively easy to predict through sequence analysis. Alignment or sliding windows can highlight residues likely to contribute to binding, distinguished by their predicted surface-exposure alongside their intrinsic chemical properties<sup>71–83</sup>.

Conformational epitopes are collections of sequentially-discontinuous residues brought into close proximity by protein folding. Most antibodies target conformational epitopes on proteins, as residues across multiple CDRs engage different regions of the antigen surface to create a more specific, complementary interface.

Attempts to predict conformational epitopes began with generic protein-protein interface prediction algorithms (see the review by Esmailbeiki *et al.*<sup>84</sup>). However, the types of contacts found in Ab-Ag complexes were soon shown to be different from those found in general protein-protein interactions<sup>85–87</sup>, displaying a unique pattern both in terms of amino acid usage and in binding site interactions<sup>70,85,86</sup>. More recent predictors seek improved performance by accounting for this specialised binding, for example by retraining existing protein-protein interaction predictors only on Ab-Ag structures<sup>88–103</sup>. While progress has been made, prediction precision is still close to random for most tools, meaning we remain unable to distinguish the region(s) of an antigen surface that are generally more prone to being bound by antibodies.

Increasingly proteins are found to have multiple epitopes, implying that many, often overlapping, surface patches on a protein antigen can engage an antibody<sup>104,105</sup>. Epitope prediction algorithms have therefore evolved to incorporate properties of the partner antibody as inputs<sup>106–112</sup>. With this new perspective, both graph-based approaches<sup>108–110</sup> and neural networks<sup>111</sup> have demonstrated an improvement in epitope prediction. This is particularly relevant to the field of Ig-seq Ab-Ag complex modelling, where we predefine a region of interest on the antigen surface (Fig. 1). In this context, we want to predict for each antibody model which antigen residues within the region

of interest are likely to constitute an epitope, were the two proteins to come together. Docking scoring functions (see Section 6) would then hopefully distinguish the few complementary partners from the many non-complementary partners.

Recently, Bourquard *et al.*<sup>113</sup> have released a pipeline that performs global molecular docking of an antibody into an antigen, and demonstrated the potential of *in silico* epitope mapping. Their algorithm, MAbTope, predicts epitope residues based on consensus epitopes shared by top-ranked poses<sup>113</sup>. This protocol is highly relevant if a number of antigen binders are known, but their respective epitopes remain unclear. However, its lack of scalability severely limits its use in high-throughput Ig-seq Ab-Ag complex modelling.

## 6 Antibody-antigen docking

With atomic representations of the antibody and antigen, docking proposes potential binding configurations of two molecular partners, by assessing surface complementarity. The first docking methods were designed for use in small molecule drug discovery, predicting protein-ligand binding interfaces<sup>114</sup>. Since 2005<sup>115</sup>, molecular docking tools have been generalised to allow macromolecular docking, enabling the prediction of protein-protein binding interfaces. Docking algorithms typically survey the conformational space for many binding pose guesses ('decoys'), and then rank them based on a scoring function to highlight the most probable, low-energy configuration(s). In recent years, improvements have been made to both sampling and scoring.

As the initial positioning of binding partners heavily biases sampling towards a particular binding site, global docking algorithms were developed to offer an unbiased sampling of potential binding sites across the antigen surface<sup>115–120</sup>. They generate coarse representations of each complex structure, followed by an evaluation of the shape and physicochemical complementarity at the interaction site. While these methods may seem appealing for Ab-Ag complex modelling, particularly given the inaccuracies in epitope prediction, their computational expense remains prohibitively high for use with large Ig-seq model datasets. To model thousands of potential complexes, we must therefore accept the sampling bias resulting from predefined paratopes (see Section 4) and epitopes (see Section 5) for computational tractability.

Many docking algorithms are 'rigid-body', meaning that both binding partners are prevented from exploring conformational degrees of freedom during pose generation. The payoff for this is that these methods are very rapid, taking advantage of fast Fourier transform algorithms. However, some binding sites are not accessible through a 'lock and key' binding analysis, and removing these conformational constraints can improve binding site identification, as well as ranking<sup>121</sup>. For each pose the backbone and side chain conformations at the interface can be optimised for interfacial energy and conformational entropy<sup>122–126</sup>. Such approaches could be especially advantageous in Ab-Ag complex modelling, as compensation could be made for CDR loops with lower predicted accuracy<sup>27</sup>, or higher predicted flexibility<sup>127</sup>. Though knowledge-based constraints<sup>122,123,125</sup> and the advent of algorithms optimised for graphical processors<sup>128</sup> have improved their efficiency, flexible docking methods remain too slow to use

on every model derived from an Ig-seq dataset. By scaling down the number of query antibodies to a more tractable size, it may, however, become possible to make use of these algorithms. This could be achieved by further pre-clustering of the Ig-seq models by paratope diversity, or by pre-screening for antibody paratopes likely to be complementary to a chosen epitope prior to docking. The latter strategy has been validated in protein-ligand complex modelling, where only ligands that possessed desired, critical pharmacophores were retained to enrich the dataset with true binders<sup>129</sup>. An alternative approach could be to first run a rapid rigid docking protocol on all Ig-seq models to eliminate the least complementary antibodies, and then run a flexible refinement process on the remaining complexes.

Bespoke Ab-Ag scoring functions have been built that take into account the unique binding tendencies of antibodies. For example, by comparing the epitope of the docked complex to its predicted likelihood of being the actual epitope, Krawczyk *et al.*<sup>110</sup> used their EpiPred score to improve the ranking of docked poses. Ramírez-Aportela *et al.*<sup>130</sup> developed uniquely-weighted scoring schemes for several classes of protein-protein complex, including Ab-Ag complexes. Through their FRODOCK algorithm, they proved that these optimised weights can improve the scores of correct complexes in each class. However, there is still considerable room for improvement, as even class-specific ranking schemes have limited success in consistently recognising the near-native decoy as the 'top hit'<sup>110,130</sup>. Until this improves, it may be advantageous to examine the properties of several top-ranking decoys.

## 7 Discussion

In this review, we have laid out a framework for discovering antibodies from Ig-seq datasets that bind to a specific epitope, without the need for prior knowledge of other binders.

Achieving this task requires a careful compromise between computational efficiency and accuracy, and we have sought to distinguish highly-refined algorithms that could be useful for small datasets of known binders, against those that are currently tractable for Ig-seq dataset analysis (see the Supplementary Information for summary tables). The resulting set of predicted binding antibodies will not solely contain genuine partners, however it should be heavily enriched for true binders.

In recent years, the need to consider developability alongside binding affinity when designing a therapeutic has become better appreciated<sup>4</sup>. While harnessing human-derived sequences is likely to minimise some developability issues, such as immunogenicity, there is no guarantee that derived antibodies will not be polyspecific, nor that they will not aggregate, be prohibitively viscous, or become insoluble when stored in vial concentrations. *In silico* developability assessment protocols ought therefore to be included in the pipeline to reduce these risks, either by removing problematic sequences or to flag them as likely to need further refinement.

With both affinity and developability in mind, large scale sequencing datasets of natively-expressed Ab chains can be filtered into an experimentally-testable number of sequences. The most promising candidates can then be further experimentally or com-

putationally refined (we recommend the review by Sormanni *et al.*<sup>4</sup> for a comprehensive description of proven refinement techniques). Case studies have already demonstrated the potential that Ab-Ag complex modelling holds in designing therapeutic antibodies against influenza and dengue viruses<sup>131–133</sup>.

Major challenges still remain in improving the accuracy of high-throughput Ab-Ag complex prediction. We have discussed the challenges related to homology modelling long CDRH3 loops, which are required to access certain epitopes<sup>62</sup>. Our low confidence in modelling accuracy for these longer loops may however be compensated with greater allowed conformational flexibility in docking. A second challenge for some antigens is a lack of good structural knowledge, which will hamper all subsequent steps.

Paratope and epitope prediction is essential for efficient docking, but both can be notoriously difficult to predict. In particular, we remain unable to accurately forecast the energetic importance of each paratopic and epitopic component to binding affinity. Finally, we described how Ab-Ag specific docking scoring functions still require significant additional development to reliably assign the best score to the most accurate decoy. Considering the characteristics of a set of top-ranked decoys for each model may be necessary to achieve more reliable results in the interim.

Despite these challenges, we believe that the exciting era of computational mAb lead generation from Ig-seq data is imminent. Successful implementation of these methods would significantly reduce the time and expense required for early-stage mAb drug discovery, making new disease targets more economically viable to investigate.

## Conflicts of Interest

There are no conflicts to declare.

## Acknowledgements

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].

## References

- 1 J. A. Weinstein, N. Jiang, R. A. I. White, D. S. Fisher and S. R. Quake, *Science*, 2009, **324**, 807–810.
- 2 A. Kovaltsuk, J. Leem, S. Kelm, J. Snowden, C. M. Deane and K. Krawczyk, *Journal of Immunology*, 2018, **201**, 2502–2509.
- 3 B. Briney, A. Inderbitzin, C. Joyce and D. R. Burton, *Nature*, 2019, **566**, 393–397.
- 4 P. Sormanni, F. A. Aprile and M. Vendruscolo, *Chemical Society Reviews*, 2018, **47**, 9137–9157.
- 5 M. I. J. Raybould, C. Marks, K. Krawczyk, B. Taddese, J. Nowak, A. P. Lewis, A. Bujotzek, J. Shi and C. M. Deane, *Proceedings of the National Academy of Sciences USA*, 2019, **116**, 4025–4030.
- 6 J. Dunbar, K. Krawczyk, J. Leem, T. Baker, A. Fuchs, G. Georges, J. Shi and C. M. Deane, *Nucleic Acids Research*, 2014, **42**, D1140–D1146.
- 7 J. B. Charbonnier, B. Golinelli-Pimpaneau, B. Gigant, D. S.

- Tawfik, R. Chap, D. G. Schindler, S. H. Kim, B. S. Green, Z. Eshhar and M. Knossow, *Science*, 1997, **275**, 1140–1142.
- 8 T. T. Wu and E. A. Kabat, *The Journal of Experimental Medicine*, 1970, **132**, 211–250.
  - 9 C. Chothia and A. M. Lesk, *Journal of Molecular Biology*, 1987, **196**, 901–917.
  - 10 R. M. MacCallum, A. C. Martin and J. M. Thornton, *Journal of Molecular Biology*, 1996, **262**, 732–745.
  - 11 B. North, A. Lehmann and R. L. Dunbrack Jr., *Journal of Molecular Biology*, 2011, **406**, 228–256.
  - 12 J. C. Almagro, A. Teplyakov, J. Luo, R. W. Sweet, S. Kodan-gattil, F. Hernandez-Guzman and G. L. Gilliland, *Proteins*, 2014, **82**, 1553–1562.
  - 13 J. Dunbar, K. Krawczyk, J. Leem, C. Marks, J. Nowak, C. Regep, G. Georges, S. Kelm, B. Popovic and C. M. Deane, *Nucleic Acids Research*, 2016, **44**, W474–W478.
  - 14 B. J. DeKosky, G. C. Ippolito, R. P. Deschner, J. J. Lavinder, Y. Wine, B. M. Rawlings, N. Varadarajan, C. Giesecke and D. T., *Nature Biotechnology*, 2014, **31**, 166–169.
  - 15 B. J. DeKosky, O. I. Lungu, D. Park, E. L. Johnson, W. Charab, C. Chrysostomou, D. Kuroda, A. D. Ellington, G. C. Ippolito, J. J. Gray and G. Georgiou, *Proceedings of the National Academy of Sciences USA*, 2016, **113**, E2636–E2645.
  - 16 P. Bannas, J. Hambach and F. Koch-Nolte, *Frontiers in Immunology*, 2017, **8**, 1603.
  - 17 L. S. Mitchell and L. J. Colwell, *Proteins*, 2018, **86**, 697–706.
  - 18 X. Li, X. Duan, K. Yang, W. Zhang, C. Zhang, L. Fu, Z. Ren, C. Wang, J. Wu, R. Lu, Y. Ye, M. He, C. Nie, N. Yang, J. Wang, H. Yang, X. Liu and W. Tan, *PLoS ONE*, 2016, **11**, e0161801.
  - 19 S. T. Reddy, X. Ge, A. E. Miklos, R. A. Hughes, S. H. Kang, K. H. Hoi, C. Chrysostomou, S. P. Hunicke-Smith, B. L. Iverson, P. W. Tucker, A. D. Ellington and G. Georgiou, *Nature Biotechnology*, 2010, **28**, 965–969.
  - 20 N. T. Gupta, K. D. Adams, A. W. Briggs, S. C. Timberlake, F. Vigneault and S. H. Kleinstein, *Journal of Immunology*, 2017, **198**, 2489–2499.
  - 21 H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov and P. E. Bourne, *Nucleic Acids Research*, 2000, **28**, 235–242.
  - 22 A. Kovaltsuk, K. Krawczyk, J. D. Galson, D. F. Kelly, C. M. Deane and J. Trück, *Frontiers in Immunology*, 2017, **8**, 1753.
  - 23 K. Krawczyk, S. Kelm, A. Kovaltsuk, J. D. Galson, D. Kelly, J. Trück, C. Regep, J. Leem, W. K. Wong, J. Nowak, J. Snowden, M. Wright, L. Starkie, A. Scott-Tucker, J. Shi and C. M. Deane, *Frontiers in Immunology*, 2018, **9**, 1698.
  - 24 W. K. Wong, G. Georges, F. Ros, S. Kelm, A. P. Lewis, B. Taddese, J. Leem and C. M. Deane, *Bioinformatics*, 2018, **34**, e0877.
  - 25 C. M. Deane and T. L. Blundell, *Protein Science*, 2001, **10**, 599–612.
  - 26 Y. Choi and C. M. Deane, *Proteins*, 2010, **78**, 1431–1440.
  - 27 J. Leem, J. Dunbar, G. Georges, J. Shi and C. M. Deane, *mAbs*, 2016, **8**, 1259–1268.
  - 28 K. Yamashita, K. Ikeda, K. Amada, S. Liang, Y. Tsuchiya, H. Nakamura, H. Shirai and D. M. Standley, *Bioinformatics*, 2014, **30**, 3279–3280.
  - 29 M. S. Klausen, M. V. Anderson, M. C. Jespersen, M. Nielsen and P. Marcatili, *Nucleic Acids Research*, 2015, **43**, W349–W355.
  - 30 P. Marcatili, A. Rosi and A. Tramontano, *Bioinformatics*, 2008, **24**, 1953–1954.
  - 31 B. D. Weitzner, J. R. Jeliazkov, S. Lyskov, N. Marze, D. Kuroda, R. Frick, J. Adolf-Bryfogle, N. Biswas, R. L. Dunbrack Jr and J. J. Gray, *Nature Protocols*, 2017, **12**, 401–416.
  - 32 H. Kemmish, M. Fasnacht and L. Yan, *PLoS One*, 2017, **12**, e0177923.
  - 33 J. K. X. Maier and P. Labute, *Proteins*, 2014, **82**, 1599–1610.
  - 34 A. Bujotzek, A. Fuchs, C. Qu, J. Benz, S. Klostermann, I. Antes and G. Georges, *mAbs*, 2015, **7**, 838–852.
  - 35 K. Zhu, T. Day, B. Warshaviak, C. Murrett, R. Friesner and D. Pearlman, *Proteins*, 2014, **82**, 1646–1655.
  - 36 M. Berrondo, S. Kaufmann and M. Berrondo, *Proteins*, 2014, **82**, 1636–1645.
  - 37 Y. Karami, F. Guyon, S. De Vries and P. Tufféry, *Scientific Reports*, 2018, **8**, 13673.
  - 38 M. A. Messih, R. Lepore and A. Tramontano, *Bioinformatics*, 2015, **31**, 3767–3772.
  - 39 P. W. Hildebrand, A. Goede, R. A. Bauer, B. Gruening, J. Ismer, E. Michalsky and R. Preissner, *Nucleic Acids Research*, 2009, **37**, W571–W574.
  - 40 N. Fernandez-Fuentes, J. Zhai and A. Fiser, *Nucleic Acids Research*, 2006, **34**, W173–W176.
  - 41 E. Michalsky, A. Goede and R. Preissner, *Protein Engineering, Design and Selection*, 2003, **16**, 979–985.
  - 42 D. Holtby, S. C. Li and M. Li, *Journal of Computational Biology*, 2013, **20**, 212–223.
  - 43 K. R. Abhinandan and A. C. R. Martin, *Protein Engineering, Design and Selection*, 2010, **23**, 689–697.
  - 44 J. Dunbar, A. Fuchs, J. Shi and C. Deane, *Protein Engineering Design and Selection*, 2013, **26**, 611–620.
  - 45 A. Bujotzek, J. Dunbar, F. Lipsmeier, W. Schäfer, I. Antes, C. M. Deane and G. Georges, *Proteins*, 2015, **83**, 681–695.
  - 46 G. G. Krivov, M. V. Shapovalov and R. L. Dunbrack Jr., *Proteins*, 2009, **11**, 778–795.
  - 47 Z. Miao, Y. Cao and T. Jiang, *Bioinformatics*, 2011, **27**, 3117–3122.
  - 48 K. Nagata, A. Randall and P. Baldi, *Proteins*, 2012, **80**, 142–153.
  - 49 C. W. Wood, M. Bruning, A. A. Ibarra, G. J. Bartlett, A. R. Thomson, R. B. Sessions, R. L. Brady and D. N. Woolfson, *Bioinformatics*, 2014, **30**, 3029–3035.
  - 50 J. Leem, G. Georges, J. Shi and C. M. Deane, *Proteins*, 2018, **86**, 383–392.
  - 51 C. Marks, J. Nowak, S. Klostermann, G. Georges, J. Dunbar, J. Shi, S. Kelm and C. M. Deane, *Bioinformatics*, 2017, **33**, 1346–1353.
  - 52 A. Stein and T. Kortemme, *PLoS One*, 2013, **8**, 1–13.
  - 53 C. S. Soto, M. Fasnacht, J. Zhu, L. Forrest and B. Honig,



- Proteins*, 2008, **70**, 834–843.
- 54 S. Liang, C. Zhang and Y. Zhou, *Journal of Computational Chemistry*, 2013, **35**, 335–341.
  - 55 M. P. Jacobsen, D. L. Pincus, C. S. Rapp, T. J. Day, B. Honig, D. E. Shaw and R. A. Friesner, *Proteins*, 2004, **55**, 351–367.
  - 56 A. Fiser, R. K. G. Do and A. Šali, *Protein Science*, 2008, **9**, 1753–1773.
  - 57 G. Wang and R. L. Dunbrack Jr, *Bioinformatics*, 2003, **19**, 1589–1591.
  - 58 M. Fasnacht, K. Butenhof, A. Goupil-Lamy, F. Hernandez-Guzman, H. Hongwei and L. Yan, *Proteins*, 2014, **82**, 1583–1598.
  - 59 A. C. Martin, J. C. Cheetham and A. R. Rees, *Proceedings of the National Academy of Sciences USA*, 1989, **86**, 9268–9272.
  - 60 N. R. J. Whitelegg and A. R. Rees, *Protein Engineering Design and Selection*, 2000, **13**, 819–824.
  - 61 C. Marks and C. M. Deane, *Computational and Structural Biotechnology Journal*, 2017, **15**, 222–231.
  - 62 I. Setliff, W. J. McDonnell, N. Raju, R. G. Bombardi, A. Murji, C. Scheepers, R. Ziki, C. Mynhardt, B. E. Shepherd, A. A. Mamchak, N. Garrett, S. A. Karim, S. A. Mallal, J. E. Crowe Jr, L. Morris and I. S. Georgiev, *Cell Host and Microbe*, 2018, **23**, P845–854.
  - 63 P. Sormanni, F. A. Aprile and M. Vendruscolo, *Journal of Molecular Biology*, 2015, **427**, 478–490.
  - 64 T. M. Lauer, N. J. Agrawal, N. Chennamsetty, K. Egodage, B. Helk and T. B. L., *Journal of Pharmaceutical Sciences*, 2012, **101**, 102–115.
  - 65 A. Jarasch, H. Koll, J. T. Regula, M. Bader, A. Papadimitriou and H. Kettenberger, *Journal of Pharmaceutical Sciences*, 2015, **104**, 1885–1898.
  - 66 V. Kunik, B. Peters and Y. Ofran, *PLoS Computational Biology*, 2012, **8**, e1002388.
  - 67 V. Kunik, S. Ashkenazi and Y. Ofran, *Nucleic Acids Research*, 2012, **40**, W521–W524.
  - 68 P. P. Olimpieri, A. Chailyan, A. Tramontano and P. Marcattili, *Bioinformatics*, 2013, **29**, 2285–2291.
  - 69 E. Liberis, P. Veličković, P. Sormanni, M. Vendruscolo and P. Liò, *Bioinformatics*, 2018, **34**, 2944–2950.
  - 70 K. Krawczyk, T. Baker, J. Shi and C. M. Deane, *Protein Engineering Design and Selection*, 2013, **26**, 621–629.
  - 71 E. A. Emini, J. V. Hughes, D. S. Perlow and J. Boger, *Journal of Virology*, 1985, **55**, 836–839.
  - 72 P. A. Karplus and G. E. Schulz, *Naturwissenschaften*, 1985, **72**, 212–213.
  - 73 J. M. Parker, D. Guo and R. S. Hodges, *Biochemistry*, 1986, **25**, 5425–5432.
  - 74 S. Saha and G. P. S. Raghava, *Artificial Immune Systems. ICARIS 2004. Lecture Notes in Computer Science*, Springer, Berlin, Heidelberg, 2004, pp. 197–204.
  - 75 I. Mayrose, T. Shlomi, N. D. Rubinstein, J. M. Gershoni, E. Ruppin, R. Sharan and T. Pupko, *Nucleic Acids Research*, 2006, **35**, 69–78.
  - 76 E. M. Bublil, N. T. Freund, I. Mayrose, O. Penn, A. Roitburd-Berman, N. D. Rubinstein, T. Pupko and J. M. Gershoni, *Proteins: Structure, Function, and Bioinformatics*, 2007, **68**, 294–304.
  - 77 Y. El-Manzalawy, D. Dobbs and V. Honavar, *Journal of Molecular Recognition*, 2008, **21**, 243–255.
  - 78 Y. El-Manzalawy, D. Dobbs and V. Honavar, *Computational Systems Bioinformatics: (Volume 7)*, World Scientific, 2008, pp. 121–132.
  - 79 M. J. Sweredoski and P. Baldi, *Protein Engineering, Design and Selection*, 2009, **22**, 113–120.
  - 80 B. Yao, L. Zhang, S. Liang and C. Zhang, *PLoS One*, 2012, **7**, e45152.
  - 81 W. Chen, W. W. Guo, Y. Huang and Z. Ma, *PLoS One*, 2012, **7**, e37869.
  - 82 J. Gao, E. Faraggi, Y. Zhou, J. Ruan and L. Kurgan, *PLoS One*, 2012, **7**, e40104.
  - 83 H. Singh, H. R. Ansari and G. P. S. Raghava, *PLoS One*, 2013, **8**, e62216.
  - 84 R. Esmailbeiki, K. Krawczyk, B. Knapp, J.-C. Nebel and C. M. Deane, *Briefings in Bioinformatics*, 2015, **17**, 117–131.
  - 85 T. Ramaraj, T. Angel, E. A. Dratz, A. J. Jesaitis and B. Mumey, *Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics*, 2012, **1824**, 520–532.
  - 86 H.-P. Peng, K. H. Lee, J.-W. Jian and A.-S. Yang, *Proceedings of the National Academy of Sciences USA*, 2014, **111**, E2656–E2665.
  - 87 D. Kuroda and J. J. Gray, *Bioinformatics*, 2016, **32**, 2451–2456.
  - 88 U. Kulkarni-Kale, S. Bhosle and A. S. Kolaskar, *Nucleic Acids Research*, 2005, **33**, W168–W171.
  - 89 A. Schreiber, M. Humbert, A. Benz and U. Dietrich, *Journal of Computational Chemistry*, 2005, **26**, 879–887.
  - 90 S. Liang, S. Liu, C. Zhang and Y. Zhou, *Proteins: Structure, Function, and Genetics*, 2007, **69**, 244–253.
  - 91 J. Ponomarenko, H. H. Bui, W. Li, N. Fusseder, P. E. Bourne, A. Sette and B. Peters, *BMC Bioinformatics*, 2008, **9**, 514.
  - 92 M. J. Sweredoski and P. Baldi, *Bioinformatics*, 2008, **24**, 1459–1460.
  - 93 S. S. Negi and W. Braun, *Bioinformatics and Biology Insights*, 2009, **3**, 71–81.
  - 94 N. D. Rubinstein, I. Mayrose, E. Martz and T. Pupko, *BMC Bioinformatics*, 2009, **10**, 287.
  - 95 J. Sun, D. Wu, T. Xu, X. Wang, X. Xu, L. Tao, Y. X. Li and Z. W. Cao, *Nucleic Acids Research*, 2009, **37**, W612–W616.
  - 96 H. R. Ansari and G. P. Raghava, *Immunome Research*, 2010, **6**, 6.
  - 97 S. Liang, D. Zheng, D. M. Standley, B. Yao, M. Zacharias and C. Zhang, *BMC Bioinformatics*, 2010, **11**, 381.
  - 98 L. F. Pacios, L. Tordesillas, A. Palacin, R. Sanchez-Monge, G. Salcedo and A. Diaz-Perales, *Journal of Chemical Information and Modeling*, 2011, **51**, 1465–1473.
  - 99 L. Zhao, L. Wong, L. Lu, S. C. Hoi and J. Li, *BMC Bioinformatics*, 2012, **13** (Suppl 17), S20.
  - 100 J. V. Kringelum, C. Lundegaard, O. Lund and M. Nielsen,



- PLoS Computational Biology*, 2012, **8**, e1002829.
- 101 G. A. Dalkas and M. Rooman, *BMC Bioinformatics*, 2017, **18**, 95.
  - 102 M. C. Jespersen, B. Peters, M. Nielsen and P. Marcatili, *Nucleic Acids Research*, 2017, **45**, W24–W29.
  - 103 L. Zhao, S. Wu, J. Jiang, W. Li, J. Luo and J. Li, *Bioinformatics*, 2018, **34**, 2061–2068.
  - 104 J. A. Greenbaum, P. H. Andersen, M. Blythe, H.-H. Bui, R. E. Cachau, J. Crowe, M. Davies, A. S. Kolaskar, O. Lund, S. Morrison, B. Mumey, Y. Ofra, J.-L. Pellequer, C. Pinilla, J. V. Ponomarenko, G. P. S. Raghava, M. H. V. van Regenmortel, E. L. Roggen, A. Sette, A. Schlessinger, J. Sollner, M. Zand and B. Peters, *Journal of Molecular Recognition: An Interdisciplinary Journal*, 2007, **20**, 75–82.
  - 105 V. Kunik and Y. Ofra, *Protein Engineering, Design and Selection*, 2013, **26**, 599–609.
  - 106 R. Rapberger, A. Lukas and B. Mayer, *Journal of Molecular Recognition: An Interdisciplinary Journal*, 2007, **20**, 113–121.
  - 107 S. Soga, D. Kuroda, H. Shirai, M. Kobori and N. Hirayama, *Protein Engineering, Design and Selection*, 2010, **23**, 441–448.
  - 108 L. Zhao and J. Li, *BMC Structural Biology*, 2010, **10**, S6.
  - 109 L. Zhao, L. Wong and J. Li, *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, 2011, **8**, 1483–1494.
  - 110 K. Krawczyk, X. Liu, T. Baker, J. Shi and C. M. Deane, *Bioinformatics*, 2014, **30**, 2288–2294.
  - 111 S. Ahmad and K. Mizuguchi, *PLoS One*, 2011, **6**, e29104.
  - 112 I. Sela-Culang, S. Ashkenazi, B. Peters and Y. Ofra, *Bioinformatics*, 2014, **31**, 1313–1315.
  - 113 T. Bourquard, A. Musnier, V. Puard, S. Tahir, M. A. Ayoub, Y. Jullian, T. Boulo, N. Gallay, H. Watier, G. Bruneau, E. Reiter, P. Crépieux and A. Poupon, *Journal of Immunology*, 2018, **201**, ji1701722.
  - 114 I. D. Kuntz, J. M. Blaney, S. J. Oatley, R. Langridge and T. E. Ferrin, *Journal of Molecular Biology*, 1982, **161**, 269–288.
  - 115 D. Schneidman-Duhovny, Y. Inbar, R. Nussinov and H. J. Wolfson, *Nucleic Acids Research*, 2005, **33**, W363–W367.
  - 116 D. Kozakov, R. Brenke, S. R. Comeau and S. Vajda, *Proteins: Structure, Function, and Bioinformatics*, 2006, **65**, 392–406.
  - 117 A. Tovchigrechko and I. A. Vakser, *Nucleic Acids Research*, 2006, **34**, W310–W314.
  - 118 B. G. Pierce, K. Wiehe, H. Hwang, B. H. Kim, T. Vreven and Z. Weng, *Bioinformatics*, 2014, **30**, 1771–1773.
  - 119 N. Shimba, N. Kamiya and H. Nakamura, *Journal of Chemical Information and Modeling*, 2016, **56**, 2005–2012.
  - 120 D. Kozakov, D. R. Hall, B. Xia, K. A. Porter, D. Padhorny, C. Yueh, D. Beglov and S. Vajda, *Nature Protocols*, 2017, **12**, 255.
  - 121 N. S. Pagadala, K. Syed and J. Tuszynski, *Biophysical Reviews*, 2017, **9**, 91–102.
  - 122 A. Sircar and J. J. Gray, *PLoS Computational Biology*, 2010, **6**, e1000644.
  - 123 M. Torchala, I. H. Moal, R. A. Chaleil, J. Fernandez-Recio and P. A. Bates, *Bioinformatics*, 2013, **29**, 807–809.
  - 124 T. Li, R. J. Pantazes and C. D. Maranas, *PLoS One*, 2014, **9**, e105954.
  - 125 G. C. P. Van Zundert, J. P. G. L. M. Rodrigues, M. Trellet, C. Schmitz, P. Kastiris, E. Karaca, A. S. J. Melquiond, M. van Dijk, S. J. De Vries and A. M. J. J. Bonvin, *Journal of Molecular Biology*, 2016, **428**, 720–725.
  - 126 N. A. Marze, S. S. Roy Burman, W. Sheffler and J. J. Gray, *Bioinformatics*, 2018, **1**, 9.
  - 127 C. Marks, J. Shi and C. M. Deane, *Bioinformatics*, 2018, **34**, 949–956.
  - 128 G. Macindoe, L. Mavridis, V. Venkatraman, M.-D. Devignes and D. W. Ritchie, *Nucleic Acids Research*, 2010, **38**, W445–W449.
  - 129 D. R. Koes and C. J. Camacho, *Journal of Chemical Information and Modeling*, 2011, **51**, 1307–1314.
  - 130 E. Ramírez-Aportela, J. R. López-Blanco and P. Chacón, *Bioinformatics*, 2016, **32**, 2386–2388.
  - 131 M. Pedotti, L. Simonelli, E. Livoti and L. Varani, *International Journal of Molecular Sciences*, 2011, **12**, 226–251.
  - 132 K. Tharakaraman, L. N. Robinson, A. Hatas, Y.-L. Chen, L. Siyue, S. Raguram, V. Sasisekharan, G. N. Wogan and R. Sasisekharan, *Proceedings of the National Academy of Sciences USA*, 2013, **110**, E1555–E1564.
  - 133 L. Simonelli, M. Pedotti, M. Beltramello, E. Livoti, L. Calzolari, F. Sallusto, A. Lanzavecchia and L. Varani, *PLoS One*, 2013, **8**, e55561.

# Supplementary Material for “Antibody-Antigen Complex Modelling in the Era of Immunoglobulin Repertoire Sequencing”

Matthew I. J. Raybould,<sup>a‡</sup> Wing Ki Wong,<sup>a‡</sup> and Charlotte M. Deane<sup>a</sup>

<sup>a</sup> *Department of Statistics, University of Oxford, 24-29 St. Giles’,  
Oxford, United Kingdom, OX1 3LB. E-mail: deane@stats.ox.ac.uk*

## ANTIBODY MODELLING TOOLS

Tool	Availability	Self-reported run time	Reference
ABodyBuilder	Free	30s	1
Kotai Antibody Builder		100min	2
LYRA		35s	3
PIGS		-	4
Rosetta Antibody		60min	5
Discovery Studio (BIOVIA)	Commercial	6min	6
MOE		30-75min	7
MoFvAb		-	8
BioLuminate (Schrodinger)		-	9
SmrtMolAntibody (Macromoltek)		30min	10

## PARATOPE PREDICTION TOOLS

Tool	General Description	Reference
Paratome	Sequence alignment method	11, 12
proABC	Random forest algorithm	13
Antibody i-Patch	Network approach based on binding likelihood	14
Parapred	Neural network algorithm	15

# EPITOPE PREDICTION TOOLS

Tool	Description/Novelty	Epitope Category	Reference
Emini <i>et al.</i> , 1985	Surface probability profiles	Linear	16
Karplus and Schulz, 1985	B-value of C-alpha atoms	Linear	17
Parker <i>et al.</i> , 1986	Hydrophilicity scale	Linear	18
BcePred	Physicochemical properties	Linear	19
ABCPred	First neural network approach	Linear	20
PepSurf	Combinatorial phage-display libraries	Linear	21
Mapitope	Peptide libraries from mAbs	Linear	22
COBEpro	Support vector machine (SVM) for epitopic propensity for each residues based on the fragment	Linear	23
BEST	SVM with predicted solvent accessibility and secondary structure	Linear	24
SVMTriP	SVM on tri-peptide similiarity and propensity scores	Linear	25
PepMapper	Affinity selected peptides derived from phage display, with adaptive distance threshold	Linear	26
LBtope	SVM on Chen's amino acid pair (AAP) propensities, Composition-Transition-Distribution (CTD) profile	Linear	27
APCPred	SVM on derived from amino acid anchoring pair composition (APC)	Linear	28
EPI-peptide designer	Peptide epitope designer. Find most frequent interface partners using graph analysis.	Linear	29
ElliPro	Calculate residue protrusion index, cluster neighbouring residues based on PI values	Linear / Conformational	30
Epitopia	Naive Bayes classifier	Linear / Conformational	31
EpiPred	Combines conformational matching of the antibody-antigen structures	Linear / Conformational	32
CEP	Accessibility of amino acids. First tool to predict conformational epitopes	Conformational	33
3DEX	Physicochemical neighborhood of C-alpha/-beta atoms	Conformational	34
EPCES	Consensus scoring of propensity and physicochemical properties	Conformational	35
Rapberger <i>et al.</i> ,2007	First antibody-specific epitope prediction tool; based on shape complementarity	Conformational	36
PEPITO	New physicochemical properties	Conformational	37
EpiSearch	Patch analysis that identifies contiguous clusters of residues on the surface of antigen with similar physical-chemical properties as found in phage display sequences	Conformational	38
SEPPA	Define clustering coefficient and residue neighbor of epitope patches	Conformational	39
CBTOPE	SVM on composition profile of patterns	Conformational	40
EPSVR	Support vector regression on physicochemical properties	Conformational	41
ASEP	Occurrence of residue pairs at epitope-paratope interface, followed by antibody-specific epitope propensity	Conformational	42
Bepar	Interacting residue pairs	Conformational	43
PPiPP	Neural network trained on interacting residue pairs	Conformational	44
LocaPep	Local search of epitope surface patches by residue clusters	Conformational	45
ABepar	Sequence conformational epitope prediction. coupling graph.	Conformational	46
DiscoTope 2.0	Log-odds ratio of the spatial neighbourhood and surface measures for epitope prediction	Conformational	47
BeTop	Cluster subgraphs on antigen	Conformational	48
PEASE	Residue pairing preference, with experimental input	Conformational	49
SEPIa	Random Forest (RF) / Gaussian Naive Bayes	Conformational	50
BepiPred-2.0	RF encoded with physicochemical properties	Conformational	51
Glep	Subgraph clustering for detection of epitope using SVM to detect surface patch	Conformational	52

## ANTIBODY DOCKING TOOLS

Tool	Description/Novelty	Reference	Specialisation
PIPER/ClusPro Server	FFT-based, Decoys as the Reference State potential	53, 54	Ab-Ag mode
surFit	Generalized Born energy and hydration energy based on accessible surface area (GBSA)	55	Ab-Ag
SnugDock/Rosetta	Simulate induced-fit mechanism by iterating through docking optimization	56, 57	Ab-Ag
PatchDock	Connolly dot surface, shape complementarity.	58	Ab-Ag mode
FRDOCK 2.0	Spherical harmonic formulation; optimised weights for Ab-Ag	59	Ab-Ag mode
GRAMM-X	FFT-based, uses a smoothed Lennard-Jones potential on a fine grid	60	Not Ab-Ag specific
HADDOCK/PRODIGY	Semi-flexible docking with biochemical/biophysical interaction data.	61	Not Ab-Ag specific
HexServer	FFT-based optimised with GPU	62	Not Ab-Ag specific
PIER	Local statistical properties of the protein surface at the level of atomic groups	63	Not Ab-Ag specific
ProPOSE	FFT-based with side-chain flexibility	64	Not Ab-Ag specific
pyDock/FTDock	FFT-based, considers electrostatics and desolvation energy	65	Not Ab-Ag specific
SIPPER	Residue desolvation based on solvent-exposed area with the propensity-based contribution of intermolecular residue pairs.	66	Not Ab-Ag specific
SwarmDock	Flexible docking, through local docking and particle swarm optimization.	67	Not Ab-Ag specific
ZDOCK	FFT-based, shape complementarity and energy	68	Not Ab-Ag specific
InterEvDock	Incorporates evolutionary information	69	Exclude Ab-Ag

- 
- [1] J. Leem, J. Dunbar, G. Georges, J. Shi and C. M. Deane, *mAbs*, 2016, **8**, 1259–1268.
- [2] K. Yamashita, K. Ikeda, K. Amada, S. Liang, Y. Tsuchiya, H. Nakamura, H. Shirai and D. M. Standley, *Bioinformatics*, 2014, **30**, 3279–3280.
- [3] M. S. Klausen, M. V. Anderson, M. C. Jespersen, M. Nielsen and P. Marcatili, *Nucleic Acids Research*, 2015, **43**, W349–W355.
- [4] P. Marcatili, A. Rosi and A. Tramontano, *Bioinformatics*, 2008, **24**, 1953–1954.
- [5] B. D. Weitzner, J. R. Jeliazkov, S. Lyskov, N. Marze, D. Kuroda, R. Frick, J. Adolf-Bryfogle, N. Biswas, R. L. Dunbrack Jr and J. J. Gray, *Nature Protocols*, 2017, **12**, 401–416.
- [6] H. Kemmish, M. Fasnacht and L. Yan, *PLoS One*, 2017, **12**, e0177923.
- [7] J. K. X. Maier and P. Labute, *Proteins*, 2014, **82**, 1599–1610.
- [8] A. Bujotzek, A. Fuchs, C. Qu, J. Benz, S. Klostermann, I. Antes and G. Georges, *mAbs*, 2015, **7**, 838–852.
- [9] K. Zhu, T. Day, B. Warshaviak, C. Murrett, R. Friesner and D. Pearlman, *Proteins*, 2014, **82**, 1646–1655.
- [10] M. Berrondo, S. Kaufmann and M. Berrondo, *Proteins*, 2014, **82**, 1636–1645.
- [11] V. Kunik, B. Peters and Y. Ofran, *PLoS Computational Biology*, 2012, **8**, e1002388.
- [12] V. Kunik, S. Ashkenazi and Y. Ofran, *Nucleic Acids Research*, 2012, **40**, W521–W524.
- [13] P. P. Olimpieri, A. Chailyan, A. Tramontano and P. Marcatili, *Bioinformatics*, 2013, **29**, 2285–2291.
- [14] K. Krawczyk, T. Baker, J. Shi and C. M. Deane, *Protein Engineering Design and Selection*, 2013, **26**, 621–629.
- [15] E. Liberis, P. Veličković, P. Sormanni, M. Vendruscolo and P. Liò, *Bioinformatics*, 2018, **34**, 2944–2950.
- [16] E. A. Emini, J. V. Hughes, D. S. Perlow and J. Boger, *Journal of Virology*, 1985, **55**, 836–839.
- [17] P. A. Karplus and G. E. Schulz, *Naturwissenschaften*, 1985, **72**, 212–213.
- [18] J. M. Parker, D. Guo and R. S. Hodges, *Biochemistry*, 1986, **25**, 5425–5432.
- [19] S. Saha and G. P. S. Raghava, *Artificial Immune Systems. ICARIS 2004. Lecture Notes in Computer Science*, Springer, Berlin, Heidelberg, 2004, pp. 197–204.
- [20] S. Saha and G. P. S. Raghava, *Proteins: Structure, Function, and Bioinformatics*, 2006, **65**, 40–48.

- [21] I. Mayrose, T. Shlomi, N. D. Rubinstein, J. M. Gershoni, E. Ruppin, R. Sharan and T. Pupko, *Nucleic Acids Research*, 2006, **35**, 69–78.
- [22] E. M. Bublil, N. T. Freund, I. Mayrose, O. Penn, A. Roitburd-Berman, N. D. Rubinstein, T. Pupko and J. M. Gershoni, *Proteins: Structure, Function, and Bioinformatics*, 2007, **68**, 294–304.
- [23] M. J. Sweredoski and P. Baldi, *Protein Engineering, Design and Selection*, 2009, **22**, 113–120.
- [24] J. Gao, E. Faraggi, Y. Zhou, J. Ruan and L. Kurgan, *PLoS One*, 2012, **7**, e40104.
- [25] B. Yao, L. Zhang, S. Liang and C. Zhang, *PLoS One*, 2012, **7**, e45152.
- [26] W. Chen, W. W. Guo, Y. Huang and Z. Ma, *PLoS One*, 2012, **7**, e37869.
- [27] H. Singh, H. R. Ansari and G. P. S. Raghava, *PLoS One*, 2013, **8**, e62216.
- [28] W. Shen, Y. Cao, L. Cha, X. Zhang, X. Ying, W. Zhang, K. Ge, W. Li and L. Zhong, *BioData Mining*, 2015, **8**, 14.
- [29] B. Viart, C. Dias-Lopes, E. Kozlova, C. F. Oliveira, C. Nguyen, G. Neshich, C. Chávez-Olortegui, F. Molina and L. F. Felicori, *Bioinformatics*, 2016, **32**, 1462–1470.
- [30] J. Ponomarenko, H. H. Bui, W. Li, N. Fusseder, P. E. Bourne, A. Sette and B. Peters, *BMC Bioinformatics*, 2008, **9**, 514.
- [31] N. D. Rubinstein, I. Mayrose, E. Martz and T. Pupko, *BMC Bioinformatics*, 2009, **10**, 287.
- [32] K. Krawczyk, X. Liu, T. Baker, J. Shi and C. M. Deane, *Bioinformatics*, 2014, **30**, 2288–2294.
- [33] U. Kulkarni-Kale, S. Bhosle and A. S. Kolaskar, *Nucleic Acids Research*, 2005, **33**, W168–W171.
- [34] A. Schreiber, M. Humbert, A. Benz and U. Dietrich, *Journal of Computational Chemistry*, 2005, **26**, 879–887.
- [35] S. Liang, S. Liu, C. Zhang and Y. Zhou, *Proteins: Structure, Function, and Genetics*, 2007, **69**, 244–253.
- [36] R. Rapberger, A. Lukas and B. Mayer, *Journal of Molecular Recognition: An Interdisciplinary Journal*, 2007, **20**, 113–121.
- [37] M. J. Sweredoski and P. Baldi, *Bioinformatics*, 2008, **24**, 1459–1460.
- [38] S. S. Negi and W. Braun, *Bioinformatics and Biology Insights*, 2009, **3**, 71–81.
- [39] J. Sun, D. Wu, T. Xu, X. Wang, X. Xu, L. Tao, Y. X. Li and Z. W. Cao, *Nucleic Acids Research*, 2009, **37**, W612–W616.
- [40] H. R. Ansari and G. P. Raghava, *Immunome Research*, 2010, **6**, 6.
- [41] S. Liang, D. Zheng, D. M. Standley, B. Yao, M. Zacharias and C. Zhang, *BMC Bioinformatics*, 2010, **11**, 381.
- [42] S. Soga, D. Kuroda, H. Shirai, M. Kobori and N. Hirayama, *Protein Engineering, Design and Selection*, 2010, **23**, 441–448.
- [43] L. Zhao and J. Li, *BMC Structural Biology*, 2010, **10**, S6.
- [44] S. Ahmad and K. Mizuguchi, *PLoS One*, 2011, **6**, e29104.
- [45] L. F. Pacios, L. Tordesillas, A. Palacin, R. Sanchez-Monge, G. Salcedo and A. Diaz-Perales, *Journal of Chemical Information and Modeling*, 2011, **51**, 1465–1473.
- [46] L. Zhao, L. Wong and J. Li, *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, 2011, **8**, 1483–1494.
- [47] J. V. Kringelum, C. Lundegaard, O. Lund and M. Nielsen, *PLoS Computational Biology*, 2012, **8**, e1002829.
- [48] L. Zhao, L. Wong, L. Lu, S. C. Hoi and J. Li, *BMC Bioinformatics*, 2012, **13** (Suppl 17), S20.
- [49] I. Sela-Culang, S. Ashkenazi, B. Peters and Y. Ofran, *Bioinformatics*, 2014, **31**, 1313–1315.
- [50] G. A. Dalkas and M. Rooman, *BMC Bioinformatics*, 2017, **18**, 95.
- [51] M. C. Jespersen, B. Peters, M. Nielsen and P. Marcantili, *Nucleic Acids Research*, 2017, **45**, W24–W29.
- [52] L. Zhao, S. Wu, J. Jiang, W. Li, J. Luo and J. Li, *Bioinformatics*, 2018, **34**, 2061–2068.
- [53] D. Kozakov, R. Brenke, S. R. Comeau and S. Vajda, *Proteins: Structure, Function, and Bioinformatics*, 2006, **65**, 392–406.
- [54] D. Kozakov, D. R. Hall, B. Xia, K. A. Porter, D. Padhorny, C. Yueh, D. Beglov and S. Vajda, *Nature Protocols*, 2017, **12**, 255.
- [55] N. Shimba, N. Kamiya and H. Nakamura, *Journal of Chemical Information and Modeling*, 2016, **56**, 2005–2012.
- [56] A. Sircar and J. J. Gray, *PLoS Computational Biology*, 2010, **6**, e1000644.
- [57] K. P. Kilambi and J. J. Gray, *Scientific Reports*, 2017, **7**, 8145.
- [58] D. Schneidman-Duhovny, Y. Inbar, R. Nussinov and H. J. Wolfson, *Nucleic Acids Research*, 2005, **33**, W363–W367.
- [59] E. Ramírez-Aportela, J. R. López-Blanco and P. Chacón, *Bioinformatics*, 2016, **32**, 2386–2388.
- [60] A. Tovchigrechko and I. A. Vakser, *Nucleic Acids Research*, 2006, **34**, W310–W314.
- [61] G. C. P. Van Zundert, J. P. G. L. M. Rodrigues, M. Trellet, C. Schmitz, P. Kastiris, E. Karaca, A. S. J. Melquiond, M. van Dijk, S. J. De Vries and A. M. J. J. Bonvin, *Journal of Molecular Biology*, 2016, **428**, 720–725.
- [62] G. Macindoe, L. Mavridis, V. Venkatraman, M.-D. Devignes and D. W. Ritchie, *Nucleic Acids Research*, 2010, **38**, W445–W449.
- [63] I. Kufareva, L. Budagyan, E. Raush, M. Totrov and R. Abagyan, *Proteins: Structure, Function, and Bioinformatics*, 2007, **67**, 400–417.
- [64] H. Hogues, F. Gaudreault, C. R. Corbeil, C. Deprez, T. Sulea and E. O. Purisima, *Journal of Chemical Theory and Computation*, 2018, **14**, 4938–4947.
- [65] T. M.-K. Cheng, T. L. Blundell and J. Fernandez-Recio, *Proteins: Structure, Function, and Bioinformatics*, 2007, **68**, 503–515.
- [66] C. Pons, D. Talavera, X. de la Cruz, M. Orozco and J. Fernandez-Recio, *Journal of Chemical Information and Modeling*, 2011, **51**, 370–377.
- [67] M. Torchala, I. H. Moal, R. A. Chaleil, J. Fernandez-Recio and P. A. Bates, *Bioinformatics*, 2013, **29**, 807–809.
- [68] B. G. Pierce, K. Wiehe, H. Hwang, B. H. Kim, T. Vreven and Z. Weng, *Bioinformatics*, 2014, **30**, 1771–1773.
- [69] C. Quignot, J. Rey, J. Yu, P. Tufféry, R. Guerois and J. Andreani, *Nucleic Acids Research*, 2018, **46**, W408–W416.