Spring 3-2-2023

DEEP LEARNING METHODS FOR PREDICTION OF AND ESCAPE FROM PROTEIN RECOGNITION

Bowen Dai
Bowen.Dai.GR@dartmouth.edu

DEEP LEARNING METHODS FOR PREDICTION OF AND ESCAPE FROM PROTEIN RECOGNITION

A Thesis
Submitted to the Faculty
in partial fulfillment of the requirements for the
degree of

Doctor of Philosophy

in

Computer science

by Bowen Dai

Guarini School of Graduate and Advanced Studies
Dartmouth College
Hanover, New Hampshire

Feb 2023

Examining Committee:

Chris Bailey-Kellogg, Chair

Charlotte Deane

Gevorg Grigoryan

Saeed Hassanpour

F. Jon Kull, Ph.D.
Dean of the Guarini School of Graduate and Advanced Studies

Abstract

Protein interactions drive diverse processes essential to living organisms, and thus numerous biomedical applications center on understanding, predicting, and designing how proteins recognize their partners. While unfortunately the number of interactions of interest still vastly exceeds the capabilities of experimental determination methods, computational methods promise to fill the gap. My thesis pursues the development and application of computational methods for several protein interaction prediction and design tasks. First, to improve protein-glycan interaction specificity prediction, I developed GlyBERT, which learns biologically relevant glycan representations encapsulating the components most important for glycan recognition within their structures. GlyBERT encodes glycans with a branched biochemical language and employs an attention-based deep language model to embed the correlation between local and global structural contexts. This approach enables the development of predictive models from limited data, supporting applications such as lectin binding prediction. Second, to improve protein-protein interaction prediction, I developed a unified geometric deep neural network, 'PInet' (Protein Interface Network), which leverages the best properties of both data- and physics-driven methods, learning and utilizing models capturing both geometrical and physicochemical molecular surface complementarity. In addition to obtaining state-of-the-art performance in predicting protein-protein interactions, PInet can serve as the backbone for other protein-protein interaction modeling tasks such as binding affinity prediction. Finally, I turned from

prediction to design, addressing two important tasks in the context of antibody-antigen recognition. The first problem is to redesign a given antigen to evade antibody recognition, e.g., to help biotherapeutics avoid pre-existing immunity or to focus vaccine responses on key portions of an antigen. The second problem is to design a panel of variants of a given antigen to use as "bait" in experimental identification of antibodies that recognize different parts of the antigen, e.g., to support classification of immune responses or to help select among different antibody candidates. I developed a geometry-based algorithm to generate variants to address these design problems, seeking to maximize utility subject to experimental constraints. During the design process, the algorithm accounts for and balances the effects of candidate mutations on antibody recognition and on antigen stability. In retrospective case studies, the algorithm demonstrated promising precision, recall, and robustness of finding good designs. This work represents the first algorithm to systematically design antigen variants for characterization and evasion of polyclonal antibody responses.

Acknowledgments

First, I would like to thank my advisor Prof. Chris Bailey-Kellogg for his guidance and encouragement throughout my Ph.D. study and research. Thank you for patiently introducing me to the field of computational biology given my limited biology background at the beginning. This helped me a lot during the early stages. Thank you for directing me solving research problems while giving me free space to think about and explore interesting topics. Thank you for teaching me how to be a good person and good scientist. I learned it simply by looking at you.

Second, I would like to thank Prof. Charlotte Deane from University of Oxford, Prof. Gevorg Grigoryan and Prof. Saeed Hassanpour from Dartmouth College for being my committee members and providing me with incredibly helpful advice. I am also grateful to Prof. Karl Griswold and Prof. Margie Ackerman from Thayer School of Dartmouth for teaching me how to gain biological insights starting as a computer scientist. I would also like to express my deep appreciation to Prof. Facundo Memoli from The Ohio State University. Thank you for helping me be prepared for Ph.D. and I learned a lot from all of our discussions.

I would also like to thank all friends and lab mates: Chao Chen, Chen Chen, Chongyang Bai, Congran Jin, Daniel Mattox, Deeptak Verma, Drew Chang, Hongliang Zhao, Haipeng Chen, Jacob Furlon, Jianfu Zhou, Mahita Jarjapu, Natalia Syzochenko, Ruibo Liu, Rui Liu, Samir choudhary, Spencer Mitchell, Srivamshi Pittala, Yongliang Fang, Yoonjoo Choi, Yeming Luo, Yujia Shentu and Yuwei Liu. I cannot make this

without your delighting collaboration, encouraging support or inspired discussions.

Finally, I want to shout out to my family. I especially thank my fiancé Ziyue Zoe Liu. Your patience, encouragement and optimism are my pillar over the past few years. I would like to thank my parents for providing me with this chance to explore the space of science. Needless to mention the companion provide by our handsome golden retriever Gali.

Chapter 1

Introduction

Section 1.1

Protein Recognition

The ability of proteins to sensitively and specifically recognize their partners (small molecules, nucleic acids, other proteins, etc.) drives diverse processes essential to living organisms, ranging from enzymatic catalysis to gene expression to intra- and inter-cellular communication. This thesis focuses in particular on two different types of partners, or "ligands": glycans and other proteins. Lectins are proteins that recognize glycans, e.g., on glycosylated proteins and lipids, and thus play important roles in numerous processes [11]. For example, microbial surface lectins bind to glycosylated proteins on a target cell's surface, helping to initiate infection [102]. On the flip side, some lectins also serve on the defense, having high affinity against glycoproteins on the envelopes of viruses such as HIV and SARS coronavirus [103], thereby inhibiting viral replication. Protein-protein interactions (PPIs) control wide-ranging molecular functions, including both transiently (e.g., propagating signals, such as insulin activating its receptor in the regulation of glucose homeostasis [124]) and more stably (e.g., protomers forming viral capsids). The finely tuned recognition of antigens (Ags)

by antibodies (Abs) is a significant special case of PPIs [145], playing crucial roles in the adaptive immune response by blocking key antigenic functions and helping drive immune-mediated clearance of pathogens [97].

To better understand and potentially manipulate such machinery, it is important to know not just whether or not a particular protein interacts with a particular potential ligand, but also how they interact [146]. To this end, numerous experimental techniques have been developed, with the most detailed information provided by protein complex determination methods such as X-ray crystallography (X-ray), Nuclear Magnetic Resonance spectroscopy (NMR), and cryo-electron microscopy (cryo-EM). The vast majority of solved structures have been based on X-ray crystallographic techniques, with NMR, though usually of lower resolution, stepping in for some smaller proteins that aren't readily crystallizable, as well as enabling studies of protein dynamics [147]. Historically, cryo-EM was limited to large proteins complexes and provided low resolution, but due to recent rapid methodological developments, it can now be used on smaller proteins and can provide resolution in many cases comparable to that of X-ray and NMR [9].

Other more or less precise experimental approaches have been employed to provide some amount of recognition information without solving the complex structure. Hydrogen–deuterium exchange relies on the natural exchange between a protein's exposed hydrogens and those in the surrounding solvent, a process that is blocked by ligand binding and can be detected by incorporating heavy atoms (i.e., deuterium) on one side or the other and assessing localized mass differences between free and bound states by mass spectrometry [187, 167, 1]. Alanine scanning assesses the effects on binding of single-point alanine substitutions, scanning position by position through the protein to evaluate which substitutions disrupt binding and thus are potential interaction sites [75, 117, 116, 94]. Electron microscopy-based methods map binding

sites by comparing sets of 2D images of the protein complex [158]. For protein-glycan interactions in particular, X-ray and NMR structure determination techniques are quite limited due to glycan structural diversity, incomplete glycosylation, crystallization difficulties, and chemical shift overlap [150]. Thus, other methods such as shotgun glycomics [151] and glycan microarrays [11] are used to determine which lectins bind to which glycans, without providing insights into where and how.

Section 1.2

Protein Recognition Prediction

The number of interactions of interest vastly exceeds the capabilities of the experimental methods described in the previous section, even the lower-resolution mapping methods and higher-throughput array methods. This is particularly the case in applications such as characterizing how millions of antibodies from repertoire sequencing [16] recognize their antigen targets. Computational methods, on the other hand, can scale to meet demands for such large-scale predictions of protein recognition.

1.2.1. Protein-Glycan Interactions

Glycans are carbohydrate-based polymers, i.e., polysaccharides, that mostly adopt branched tree structures with rare cases of ring structures. Protein glycosylation, the attachment of a glycan to a protein, is an important post-translational modification, with the resulting glycoproteins central many diverse biological processes. For example, the recognition of glycans on viral glycoproteins such as HIV gp120 and influenza hemagglutinin by cell surface receptors enables these viruses to enter the cell [92].

Unfortunately, predicting protein-glycan interactions is extremely hard due to the diversity in glycan structure, monosaccharide composition, and glycosidic bond arrangements [106, 22]. Current computational methods for predicting protein-glycan

interaction mainly focus on binding specificity, i.e., whether a glycan and a protein bind or not. These methods are dominated by pattern-based methods which identify common structural motifs among glycans with shared binding properties [61, 29, 82, 22]; for example the motif N-acetyl-lactosamine can be used to characterize immunogenicity. One state of the art prediction method [22] seeks to regress interaction strength by glycan motif fingerprints and was able to train a model for mammalian glycans and 352 glycan-binding proteins. A key strength of motif-based methods is their high interpretability, but this is balanced by a limited ability to represent complex patterns, e.g., current motif representations mainly consist of simple patterns defined by connected and nearby monosaccharides [29, 128]. To capture a richer representation of glycan composition, recent methods have employed language models [14] and graph convolutional networks [20], thereby attaining improved prediction performance and an ability to accurately predict host taxonomy and glycan immunogenicity. However, they still fall short of learning global structural contexts due to the limitations in their network architectures. For example, the context in which a motif is presented, i.e., the global structure surrounding the motif and the linkages connecting the motif, also has an impact on interactions [91]; for example, Neu5Acα(2-6)Gal binds differently to hemagglutinins when presented in symmetric and asymmetric glycans [171]. The need to capture structural contexts mediating recognition motivated the work in this thesis on constructing a structured latent space for glycans, learning an embedding for motifs and their contexts that enables protein-glycan interaction prediction from limited data.

1.2.2. Protein-Protein Interactions

Current computational methods for predicting PPIs can be roughly divided into physically based and data-driven approaches.

Physically based methods rely on structural modeling, e.g., via docking meth-

ods [83, 23, 177, 148, 37], to generate and evaluate possible "poses" (rotations and translations) of the proteins leading to a possible interaction. Beyond the computational cost of sampling a comprehensive set of poses, a key difficulty is scoring them. Scoring functions include carefully handcrafted potential functions [37, 83] as well as learned models, e.g., via 3D convolutional neural networks [170, 134]. Successful docking methods [177, 37, 83] typically include near-native pose in the top 5 or 10 conformations, but it remains a problem to identify exactly which one, as shown in a recent Critical Assessment of Prediction of Interactions (CAPRI) assessment [89].

In contrast, data-driven models do not attempt to directly model the interaction structure, but instead use existing data to learn patterns, based on representations of the proteins' sequences or structures, that enable prediction of the interfaces mediating recognition. While some methods [152, 43] work from sequence alone, recognition is largely dependent on three-dimensional structure [183], and thus structure-based methods can learn richer representations and attain better performance when structures are available [111]. Furthermore, while some methods are "partner independent" [62, 51], i.e., seek to predict spots on one protein that are generally amenable to recognition by any partner, "partner dependent" methods [136, 126, 84, 142, 73, 2, 137] are able to leverage information from both parties in the interaction in order to make more specific predictions in the scenario where both are known. For example, [2] trained a support vector machine (SVM) to classify whether or not a pair of residues will interact, and then post-processed this contact map prediction to yield predicted binding interfaces. [137] improved this general idea by using XGBoost classifiers [24], attaining SOA performance at the time. With the recent advance in deep learning, methods using handcrafted structural features with simple Neural Networks [73, 181], Graph Convolutional Networks (GCNs) [126, 156], or Geometric Deep Learning [51] have been developed and shown to outperform traditional machine learning methods

[136, 2, 137].

Recent approaches combine physical modeling and data-driven modeling. For example, Alphafold2-multimer [45] and RoseTTAFold [6] are based on deep modeling methods originally designed for protein structure prediction, but then extended to predict interactions. The network architectures and overall processing employed by these methods draw heavily from physical modeling approaches, but the methods use data to learn underlying parameters such as residue sequence/structure correlations. Some recent studies [52, 154] utilize SE(3)-equivariant inputs to reduce the complexity due to the input orientation, and then use geometric deep learning techniques to improve protein rigid docking, achieving close to SOA performance much faster. These results demonstrate the potential for solving complex problems such as rigid and even non-rigid docking by employing a combination of physical modeling and data-driven approaches.

As discussed above, Ab-Ag recognition is one particularly important special case of protein-protein recognition. The prediction task is often formulated in terms of predicting the "epitope", the residues on the Ag that are bound by the Ab. There are numerous approaches to epitope prediction, using Ab-Ag specific post-docking scoring [83, 15], statistical inference [85], geometric matching and machine learning scoring [84], and GCNs with attention mechanisms [126]. While the SOA performance achieved by these approaches has continually improved, the precision and/or recall still remains moderate. This motivated the work described in this thesis on combining the advantages of physically-based and data-driven methods to develop a robust model which is suitable for epitope prediction and even general protein-protein interaction binding interface prediction.

Section 1.3

Antibody Recognition Evasion

The adaptive immune response against a pathogen often culminates with the development of B cells selected and matured for potent recognition of pathogen antigens [143]. The "goal" of this humoral immune response could be characterized as developing antibodies with high affinity and specificity against the antigen, ultimately leading to beneficial functions such as neutralizing the pathogen or mediating clearance. In contrast, a "goal" of the pathogen would be to escape this recognition. Likewise, in various therapeutic applications, protein engineers can seek to modify a protein so as to disrupt Ab-Ag binding and thereby escape recognition. Techniques to evade antibody recognition have been successfully employed, for example, in the development of protein therapeutics that attain improved efficacy by escaping recognition by existing antidrug antibodies (ADAs) that developed from previous therapeutic application or natural environmental exposure [138]. Antibody recognition evasion has also been used in designing vaccine antigens to focus the immune response against an epitope of interest, by escaping responses against undesired epitopes [182]. In addition to silencing unwanted binding, antibody recognition evasion has also been used to learn binding information such as localizing antibody epitopes [118, 139, 59, 66, 98] by identifying mutations that disrupt binding and to identify binding antibodies [173] by sorting antibodies against different variants.

1.3.1. Motivating applications

The work in this thesis on antibody recognition evasion was inspired by two prospective applications we are pursuing in collaboration with experimental groups. While these collaborative efforts are beyond the scope of this thesis, their goals drove the approaches developed here, and they provide concrete examples from which we gen-

eralized our methodology.

The first motivating example is B cell epitope deletion, i.e., evasion of pre-existing antibodies, for the enzyme lysostaphin, a potent antimicrobial agent against *Staphylococcus aureus*. Since lysostaphin is itself of bacterial origin, it elicits an immune response, including anti-drug antibodies that can lead to loss of efficacy among other detrimental outcomes [132]. Recent work has led to the development of lysostaphin variants from which T cell epitopes have been depleted [185, 184], so that in theory subjects who have never been exposed to lysostaphin would not mount an immune response against it. However, since lysostaphin is derived from *Staphylococcus simulans*, another common microbe, some subjects may have already encountered it and thus not be immunologically naive, and in fact already have developed anti-lysostpnin ADAs. Importantly, underlying the antibody response is an uncharacterized, and perhaps subject-dependent, polyclonal mixture of B cells — different antibodies may recognize different epitopes on lysostaphin, and these specificities are unknown. Thus, our collaborative goal is to find functional lysostaphin variants which evade uncharacterized pre-existing B cell responses.

The other motivating example is understanding how antibodies bind to SARS-CoV2 spike protein (S protein) receptor binding domain (RBD). Unfortunately a vaccine developed against one variant may elicit less potent neutralization against newly evolved variants [54]. Thus it is helpful to understand which antibodies in the human repertoire bind to what parts of RBD [178], as this knowledge may enable the development of better vaccines able to broadly block the attachment and entry of different viral variants. We here define this type of problem as "epitope map deconvolution"; the goal is to design a panel of variants for an Ag target that will differentially disrupt Ab binding against different epitopes on the Ag. Such a panel can enable experimental localization of Abs to their epitopes according to which

variants they no longer bind. It can also serve as the basis for reagents able to sort of Abs from polyclonal serum according to their specificities [25]. We aim to develop a limited-size panel (i.e., not requiring excessive experimental effort) that efficiently and effectively uses different sets of mutations on different parts of the Ag surface to sufficiently disrupt Ag binding by a wide range of (unknown) Abs in a mixture, and thereby deconvolve and localize the Abs according to their different epitopes.

Though B cell epitope deletion and epitope map deconvolution have different goals, they share a common underlying approach: designing sets of mutations on an Ag's surface to disrupt Ab binding. For B cell epitope deletion, disruption is the goal, finding an Ag variant that escapes all Abs. For epitope map deconvolution, disruption is a means to the goal, namely identifying epitopes and the Abs that recognize them. Note that for B cell epitope deletion, we seek a single Ag variant that escapes all Abs, while for epitope map deconvolution, we seek a panel of Ag variants that work together to map and sort.

1.3.2. Previous work: B cell epitope deletion

In general, B cell epitope deletion seeks to modify the surface of the target protein, so that it no longer presents the epitopes recognized by preexisting Abs, but it retains its structure and function [57, 109]. Since alanine mutations are generally relatively benign from a stability and function standpoint, but at the same time can disrupt Ab binding, [94] used alanine scanning mutagenesis as the basis for engineering a 38-kDa fragment of *Pseudomonas* exotoxin A (PE38) to evade recognition by existing Abs. Expanding the mutational choices from alanine to also include glycine and serine, [139] engineered an Ab-evading variant of truncated diphtheria-toxin (DT) by mutating hydrophilic surface residues to amino acids randomly chosen from these three possibilities. Further enlarging the amino acid repertoire, [59] used random mutagenesis to reduce the interaction of Abs against Chemotaxis inhibitory protein of

Staphylococcus aureus. To modify undesired epitopes on the Dengue virus envelope protein domain III (DENV DIII), [133] used a combinatorial approach mutagenizing selected epitope residues to all possible amino acids. Finally, in order to better ensure maintenance of stability and function in a case study application of engineering enhanced green fluorescent protein (EGFP) to evade recognition by a particular nanobody, [27] employed a Pareto optimization approach, computationally designing mutational variants according to predicted trade-offs between disruptiveness of Ab binding and preservation of Ag stability.

1.3.3. Previous work: Partial epitope map deconvolution

As previously discussed, alanine scanning mutagenesis can be leveraged to epitope map an Ag, mutating residues to alanine and evaluating effects on antibody binding and thereby identifying which are important to recognition. In order to epitope map PE38 for a set of mouse Abs, [118] evaluated both alanine and glycine mutations at 41 highly exposed surface residues, evenly spread out over the surface. [25] designed a panel of gp120 single mutation variants to enable epitope mapping of Abs against the important HIV-1 glycoprotein 120 (gp120) CD4-binding site (CD4bs) from polyclonal serum. [66] localized epitopes of tumor Ag B7H6 against two Abs TZ47 and PB11 using a set of 6 Ag triple mutants designed by docking models. [98] further built on this approach to computationally cluster repebodies (leucine-rich repeat binding scaffold) targeting interleukin-6 (IL-6) based on shared predicted epitope specificities, and then designed a set of triple-mutant variants to experimentally validate the epitopes.

Mutating surface residues of an Ag can also yield reagents useful in isolating or sorting Abs from a polyclonal mixture. For example, to find new HIV broadly neutralizing antibodies (bnAbs), [173] resurfaced gp120 to obtain a variant that preserved the CD4bs antigenic area but eliminated all other antigenic areas. In addition to epi-

tope mapping as discussed above, [25]'s panel also enabled sorting of HIV antibodies by epitope specificity, according to "signatures" of which mutations disrupted binding. [17] used mutagenesis data of herpes simplex virus (HSV) glycoprotein D (gD) together with epitope binning experiments and docking methods in order identified four clusters of anti-gD Abs.

Section 1.4

Specific Aims

This thesis uses deep learning to improve prediction of protein recognition, for both lectin-glycan interactions and Ab-Ag interactions, and it further leverages deep learning models to design variants so as to escape Ab recognition. For lectin-glycan interactions, we designed a high dimensional representation that incorporates both local [82] and global structural [91] information and enables improved prediction of lection-glycan binding. For Ab-Ag interactions, we designed a deep learning architecture that is able to learn and leverage geometrical and physicochemical information. For Ab recognition evasion, we designed computational methods to identify beneficial mutations on an Ag predicted to reduce Ab binding while retaining protein stability, supporting both B cell epitope deletion and epitope map deconvolution.

Aim 1. Physicochemically-oriented geometric deep learning for partner-specific protein-protein interaction binding interface prediction. In order to improve protein interaction interface prediction, we leveraged the advantage of both machine learning and physics-driven methods to develop a unified geometric deep neural network, "PInet" (Protein Interface Network) [33]. In order to learn shape and physicochemical complementarity, PInet takes as input pairs of protein surface meshes and feeds them to a geometric deep learning framework that segments

their structural regions mediating interaction. We tested PInet on representative PPI datasets, matching or outperforming SOA methods on benchmark datasets from Docking benchmark Dataset v5 (DBD5) [168], EpiPred [84], and MaSIF [51].

Aim 2. Context-aware glycan embedding for lectin-glycan interaction specificity prediction. In order to improve lectin-glycan interaction specificity prediction, we developed GlyBert [34] to capture complex patterns including global contexts and local motifs. To reflect both global contexts and local motifs in a straightforward representation, we developed a tree-based glycan encoding that properly annotates the linearity and flexibility of the structure. To leverage the advantages of self-attention in learning local and global correlations, as demonstrated in natural language processing applications [93], we pre-trained a BERT [39] like language model to learn embeddings of whole glycans and their constituent monosaccharides, based on a set of semi-supervised and supervised tasks. We then fine-tuned the network on a lectin-glycan binding specificity prediction task, yielding significant improvement in predictive performance from that obtained by a motif-based method [29] and a graph neural network based method [22]. We also explored the potential of the model for design, developing a proof-of-concept generative design algorithm in which we modified glycans to transform them from non-immunogenic to immunogenic.

Aim 3. Antigen variant design for B cell epitope deletion and epitope map deconvolution. We developed algorithms to optimize Ag variants so as to disrupt Ab binding while preserving Ag stability. For B cell epitope deletion, a geometry-based algorithm designs a set of variants to disrupt potential epitopes; these are subsequently ranked by predicted effects on Ag stability and coverage of predicted epitopes. For epitope map deconvolution, a related geometry-based algorithm generates a set of variant panels, each composed of a set of variants targeting different

potential epitopes; the panels are ranked by predicted overall variant stability and epitope coverage. Retrospective case studies showed that the B cell epitope deletion algorithm can attain high recall in finding stable but disruptive mutations, while the epitope map deconvolution algorithm promises to efficiently and effectively deconvolve and localize Abs to different epitope regions.

Chapter 2

Protein Interaction Interface Region Prediction by Geometric Deep Learning

Motivation: Protein-protein interactions drive wide-ranging molecular processes, and characterizing at the atomic level *how* proteins interact (beyond just the fact *that* they interact) can provide key insights into understanding and controlling this machinery. Unfortunately, experimental determination of three-dimensional protein complex structures remains difficult and does not scale to the increasingly large sets of proteins whose interactions are of interest. Computational methods are thus required to meet the demands of large-scale, high-throughput prediction of how proteins interact, but unfortunately both physical modeling and machine learning methods suffer from poor precision and/or recall.

Results: In order to improve performance in predicting protein interaction interfaces, we leverage the best properties of both data- and physics-driven methods to develop a unified Geometric Deep Neural Network, "PInet" (Protein Interface Network). PInet consumes pairs of point clouds encoding the structures of two partner proteins, in

order to predict their structural regions mediating interaction. To make such predictions, PInet learns and utilizes models capturing both geometrical and physicochemical molecular surface complementarity. In application to a set of benchmarks, PInet simultaneously predicts the interface regions on both interacting proteins, achieving performance equivalent to or even much better than the state-of-the-art predictor for each dataset. Furthermore, since PInet is based on joint segmentation of a representation of a protein surfaces, its predictions are meaningful in terms of the underlying physical complementarity driving molecular recognition.

Availability: PInet [33] is published on Bioinformatics. PInet scripts and models are available at `https://github.com/FTD007/PInet`.

Section 2.1

Introduction

Due to the importance of protein-protein interactions in driving cellular machinery, numerous experimental and computational techniques have been developed to identify putative partners [146]. While these methods yield information about *which* pairs of proteins might interact, they don't characterize *how* they interact (Figure 2.1). Further experimental investigations or computational analyses are then necessary to determine or predict binding modes, provide mechanistic insights, and guide subsequent efforts to, e.g., design mutations to change binding affinity or specificity or identify small molecule inhibitors of an interaction. Likewise, recent advances in repertoire sequencing have enabled the collection of millions or billions of antibody sequences from different individuals and conditions [16], and promise to provide valuable insights into vaccination and natural infection, especially if how the sequenced antibodies recognize their antigen targets could also be characterized. The same

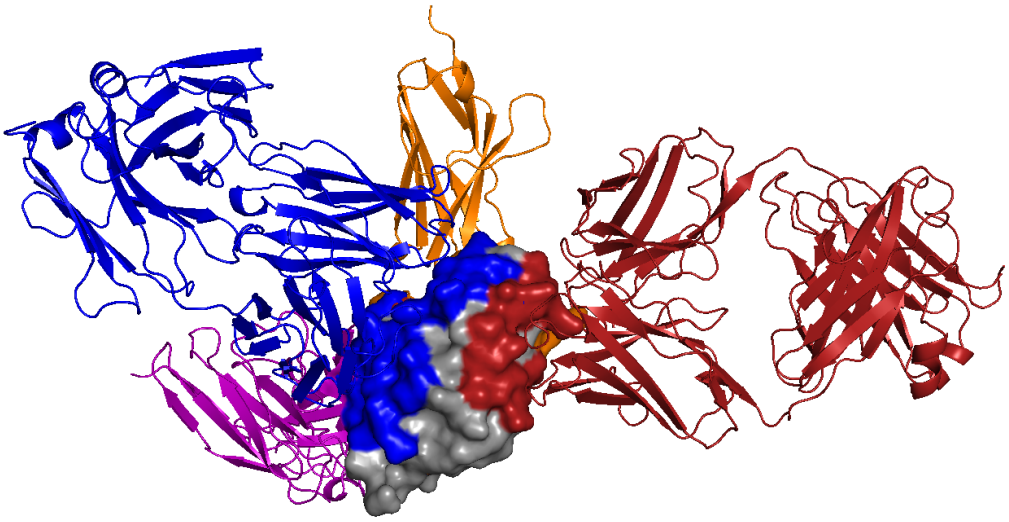

Figure 2.1: Different proteins can recognize the same protein partner in very different ways, as shown here for hen egg lysozyme (HEL; gray surface, 3LZT) and different antibodies (colored cartoon; (purple: 1BVK, blue: 1DQJ, red: 1MLC, and orange: 2I25). Partner-independent predictions seek to label, in general, what parts of one protein might be recognized by unknown other proteins. In this example, even with just these few example antibodies, much of the HEL surface is recognized by one antibody or another, so partner-independent prediction of binding region would largely cover the surface. Thus when information about particular partners is available, partner-specific predictions can be beneficial, providing a more specific characterization of recognition by separately localizing each partner.

holds for display-based development of antibody therapeutics [46], where different antibodies may have different modes of interacting with an antigen that may manifest different trade offs, and thus early characterization could drive better selection of leads. Unfortunately, while experimental structure determination provides "gold standard" insights into recognition, these and even alternative less precise / confident methods (e.g., alanine scanning) cannot keep up with the scale of experimental discovery of interacting partners.

Computational methods to predict how two given proteins interact can be roughly split into those methods based on physical models and those leveraging data-driven models. Physically-based approaches include protein-protein docking, e.g., [30, 125,

141, 172, 176], wherein structures or models of the individual partners are computationally rotated and translated (and in some cases, modified) to generate possible "poses" in which they interact. Beyond the computational cost of sampling or computing a sufficient set of poses, a key difficulty is scoring them. Scoring functions are often carefully crafted to integrate multiple factors, with shape complementarity as the foundation [88]. Typically, as has been studied with a state-of-the art physically-based docking program, ClusPro, near-native structures are generally included among the top poses, but it remains difficult to identify which one(s) [159].

In contrast to physically-based approaches, data-driven / machine learning approaches seek to leverage existing databases of characterized interactions to learn properties underlying recognition. While some methods are based on sequence alone [127, 107], structural information can be extremely important and has been shown to yield more accurate predictions [183], since recognition is largely based on surface regions that may not be contiguous in primary sequence, and as discussed above overall surface complementarity is critical. Fortunately, in many important cases, structures of the individual proteins are already available, or high-quality homology models can be readily obtained, so a number of structure-based data-driven predictors have emerged. The impact of using structural information instead of sequence is demonstrated, for example, by BIPSPI [137] and PAIRPred [2], which gain on average a 10% improvement in precision and recall using structure vs. sequence alone.

Machine learning approaches can further be classified into those that seek to predict which residues are in the interface regions, which we call here "interface region prediction", and those that seek to predict which pairs of residues (one from each partner) are interacting, which we call here "contact prediction". For example, MASIF [50] and PECAN [126] predict interface regions, ComplexContact [180] and SASNet [156] predict contacts, and the BIPSPI and PAIRPred methods mentioned above pre-

dict contacts and postprocess the results to predict interface regions. As discussed above, antibody recognition of cognate antigens is one particularly important special case of interface region prediction (the "epitope" on the antigen and the "paratope" on the antibody are the interaction regions). Epitope prediction has been the subject of much study, with DiscoTope [85] focusing on learning general surface properties, EpiPred [84] combining conformational matching and a machine learning boosted scoring function, and PECAN [126] employing graph convolution networks. Paratope prediction is also well studied, though it is somewhat easier, due to the regularity of immunoglobulin sequence and structure [126].

Different interface region prediction tasks leverage different available information to achieve different goals. *Partner-independent* prediction seeks to predict, in general, what portions of the surface of a protein may serve as interface regions for other proteins (known or unknown). In contrast, *partner-specific* prediction accounts for a particular partner in identifying the binding regions most suitable for that partner. As illustrated in Figure 2.1, deconvolving the surface into partner-specific predictions for different partners provides a better characterization of the recognition; this can be important for subsequent engineering, for understanding underlying immune responses, and so forth. In the specific case of antibodies, while many classic epitope predictors are partner-independent, [144] motivated the paradigm of antibody-specific approach, since antibodies may be developed by the immune system against a variety of different epitopes on an antigen (Figure 2.1). Further studies [142, 66] have demonstrated the utility of this framework when information about the partner antibody is available.

To bring deep learning to bear in developing new structure-based interface region prediction methods, a central question is how to represent the protein structures and thus develop a suitable neural network. Since geometry is one of the key principles

underlying interface complementarity (as exploited by physical modeling methods), but protein structures are not regularly-sampled grids like the images studied by traditional deep learning, this task belongs to the general area of geometric deep learning. Graph Convolutional Networks (GCNs) [80] are one geometric deep learning method that has been leveraged for protein interface region prediction. GCNs generalize Convolutional Neural Networks (CNNs) from 2D grids to graphs by employing spectral convolution, enabling them to avoid the unnecessary and potentially costly direct Euclidean domain representations of structured data (e.g., 3D voxelization might waste a great deal of memory at high resolution, or might lose important features at low resolution). A recent study [49] employed a GCN to learn a representation for each residue in a protein and used that representation to classify whether or not two residues interact. This study showed that convolution is able to capture interaction information from basic physicochemical properties of residues. Recently, PECAN [126] combined the advantages of GCNs and attention mechanisms [7] in an integrated model for predicting both epitopes and paratopes, learning better representations of residues and their interaction preferences in order to focus predictions on complementary regions.

Another recent protein interface region prediction method, MaSIF [50], used a different form of geometric deep learning to learn and utilize geometric features, mapping 3D surface patches of an input protein to 2D using a soft polar coordinate system, and then using CNNs to predict the likelihood of a surface vertex being involved in an interaction region. MaSIF was also trained on a substantially larger data set than previous studies, using thousands rather than hundreds of complex structures. We note that in contrast to the methods above, as well as ours, MaSIF is partner-independent, predicting likely binding sites in general for a protein, rather than making predictions that are specific to given partner proteins.

In order to directly encode and exploit surface geometry and interface complementarity in a deep learning framework, we pursue here a point cloud based representation. Traditional ways of dealing with point clouds, e.g., aggregating points into discrete voxels for a 3D voxel CNN [99], or sorting them into a linear sequence for an RNN [166], can ruin the detailed geometry of the data or result in an unstable ordering of the inherently non-linear set of points into a sequence. PointNet [131] was developed to directly learn geometrical information from point clouds, overcoming these problems by learning functions to make points order invariant, as well as aggregating local and global features over the cloud. More specifically, PointNet utilizes a Spatial Transformer Network [70] to render the input point cloud invariant to geometric transformations, employs a multi-layer perceptron to learn high-dimensional local feature vectors for the points, and subsequently applies a max pooling layer over each channel to produce global feature vectors describing overall shape features. This global feature vector can be concatenated with local feature vectors to enable semantic segmentation by learning which subset of global features should be assigned to a point. When different points share similar global signatures, their neighborhood information is also extracted, thereby helping group points for segmentation. PointNet was shown to achieve state of the art performance on problems including 3D object classification, part segmentation, and scene semantic segmentation.

Our contribution. In order to leverage the advantages of both physically-based modeling and data-driven modeling, we develop a partner-specific geometric deep learning approach to interface region prediction that is based on an explicit representation of a pair of molecular surfaces. Our approach thereby enables characterization of shape and physicochemical complementarity driving molecular recognition, using existing data to learn how best to score this complementarity and thereby identify interface regions of a given pair of structures. In addition to predicting interface regions

in general protein-protein pairs, we also address the specific case of epitope-paratope prediction in antibody-antigen (Ab-Ag) recognition. Our approach, PInet, achieves state of the art performance on each of the different interface region prediction tasks on which we evaluate it. Strikingly, even when trained on a dataset largely comprised of other types of protein-protein interactions rather than one focused specifically on antibody-antigen interactions, PInet performs better than state-of-the-art epitope predictors, demonstrating that it has learned generalized representations of protein interface complementarity.

Section 2.2

Methods

2.2.1. Problem Setup

Given individual structures or high-quality homology models of two proteins, traditionally termed the "ligand" and "receptor" (the distinction is not important in our approach), our goal is to predict their interface regions, i.e., the portions that are contacting each other. While ultimately the predicted interface regions will be characterized in terms of the involved residues, in order to directly represent recognition in terms of two complementary molecular surfaces, we encode both proteins as surface point clouds $P^l = \{P_i^l | i = 1 \dots n_l\}$ and $P^r = \{P_i^r | i = 1 \dots n_r\}$, where each point cloud includes both (x, y, z) coordinates (G_i) along with physicochemical properties (H_i, E_i) described below. We then seek to label each point as being in the interface or not, yielding an $n_l + n_r$ by 1 output probability score. The point cloud representation enables simultaneously capturing both geometric and physicochemical properties mediating recognition, and yields meaningful predictions in terms of segmented surface regions potentially mediating interactions.

2.2.2. Data preparation

Geometry. We first preprocess input PDB [10] files using PDB2PQR [40] to remove solvent molecules and fill in missing atoms. We then generate surface meshes for each protein separately and use the mesh vertices as the input point clouds; results presented here are based on PyMOL-generated meshes [38] with a water probe radius of 1.4 Å. Point coordinates in G_i are translated so that the point cloud's centroid is at the origin.

Physicochemical features. PInet allows physicochemical properties to be associated with each point. For the results shown here, we employed representative encodings of the two most important classes of properties: electrostatics (E_i) and hydrophobicity (H_i). Combined with the 3D geometry G_i, these features make P^l and P^r 5 dimensional point sets.

- **Electrostatics** Poisson–Boltzmann electrostatics are computed by APBS [8] for both proteins separately, and the continuum electrostatics value for each point (E_i) is taken as the value of the voxel containing that point normalized by the maximum value over the whole grid.

- **Hydrophobicity.** The hydrophobicity value for each point (H_i) is computed as a distance-normalized weighted sum of the Kyte-Doolittle scale [86] values for the amino acid types of the three closest residues (R_k):

$$H_i = \sum_{k=1}^{3} \mathrm{KD}(R_k) \times \frac{1}{D(P_i, R_k)}$$

  where $\mathrm{KD}(R_k)$ is the Kyte-Doolittle value for residue type R_k and $D(P_i, R_k)$ is distance between the surface point and the centroid of the residue.

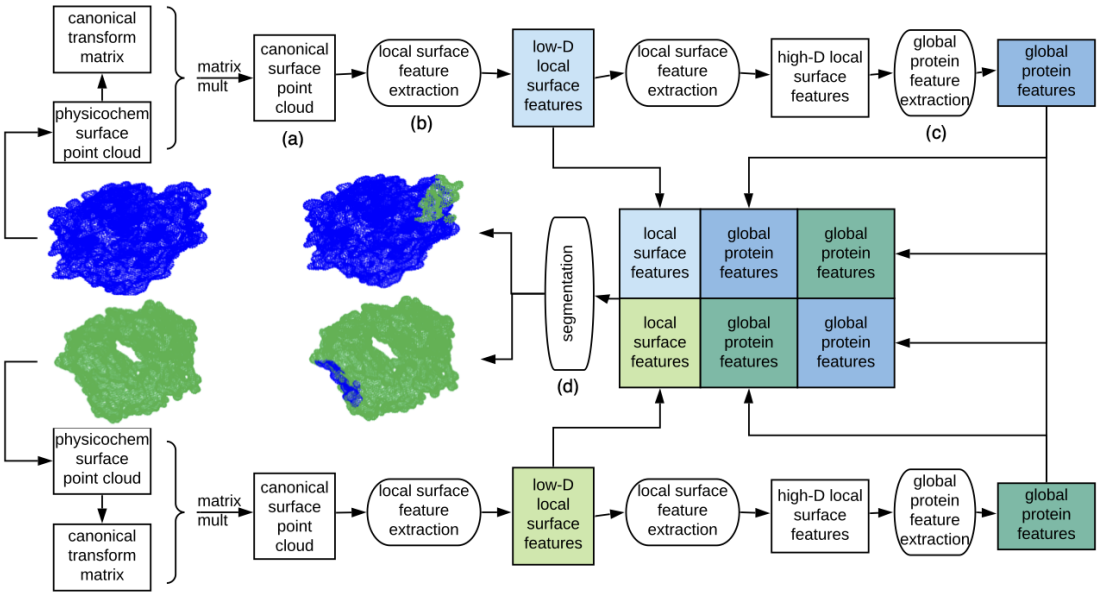

Figure 2.2: Overview of our Protein Interface Network (PInet) approach for predicting interaction regions on pairs of proteins. PInet consumes two 5 dimensional point clouds representing geometry and physicochemical properties of each protein surface, and performs a semantic segmentation on all points from both point clouds simultaneously. It first processes each point cloud separately. For each protein, a Spatial Transformation network renders the surface point clouds invariant to rigid-body transformations. Then a multi-layer perceptron (MLP) extracts local surface features. These local surface features are then aggregated into a global protein feature vector. With each protein thus processed, the protein local surface features and global protein features from both proteins are concatenated in order to be segmented by another MLP. The trainable weights for canonical transformation, local and global feature extraction are shared for the two proteins.

Interface labels. The label sets L^l and L^r assign each point a value of 1 if it is in the interface region and 0 if it isn't. Here, we define interface points as those within 2Å of a point on the partner protein (known in training from the complex structure). The 2Å threshold was selected based on an analysis of the distribution of the shortest pairwise distances across interacting proteins (Supplementary Figure 1).

2.2.3. Architecture

We use this representation as the basis for a geometric deep learning framework (Fig. 2.2) that seeks to segment points in the interface region from the rest for both input point clouds simultaneously. The architecture is rooted in that pioneered by PointNet [131]. While PointNet was developed for semantic segmentation of a single point cloud in an image according to its own features, our goal here is to learn to simultaneously segment two point clouds into interacting and non-interacting regions. Whether or not regions interact is revealed by shape and physicochemical complementarity of their sub-regions, i.e., segmentation of a set of points also depends on features of potential partner points. We thus adapt the architecture to take two point clouds as input and learn to extract feature signatures that capture complementarity between sub-regions of the clouds and enable simultaneous segmentation of both clouds into interface vs. non-interface.

Point Cloud Canonical Transformation (Fig.2.2(a)) . The first step transforms the points in a cloud into a canonical space in a way that is invariant to linear transformations. The underlying problem is that there is no standard coordinate system for protein structures, and indeed the key problem in physically-based methods is to find the right transformation (rotation and translation) to bring the two separate proteins together. Rather than sampling different poses we use a canonical representation approach, and in fact learn the transformation into the canonical space rather than hand-crafting a canonical representation. In particular, we use a simplified Spatial Transformation network [70] to learn a $k \times k$ transformation matrix (where k is the dimension of the point cloud, 5 in our case) that maps each point into the new space.

Local Surface Feature Extraction (Fig.2.2(b)) . The next step moves beyond individual points in a cloud to a richer representation of properties of the protein surface. This local surface feature extraction block is implemented by a sequence of fully connected layers on points, forming a Multilayer perceptron (MLP). Here the MLP is learning a set of functions that map input surface point features (3D geometry and 2D physicochemical) to a higher-dimensional vector of protein surface features. Inspection of the first fully connected layer of a trained model from the results reveals that some weights focus on the coordinates, aggregating geometry-related features, while other weights focus more on electrostatics and hydrophobicity channels, capturing physicochemical properties.

Global Protein Feature Extraction (Fig.2.2(c)) . After re-representing each point in terms of local surface features, the next step summarizes the entire protein surface in order to later support segmentation via comparison of local features to this global feature. The global feature is obtained by max pooling over all points' local features. This max pooling layer reveals points with dominant impact on each feature, and uses the associated features of these points to summarize the whole protein surface.

Segmentation (Fig.2.2(d)) . Finally, the binding interfaces for both proteins are simultaneously segmented. Our segmentation block is also implemented by an MLP, taking as input a combination, over both proteins, of all the local surface features and the two global feature vectors. The MLP learns complementarity from the two global feature vectors, and uses the relationship between the complementarity signature and the local feature vectors in order to predict for each point its probability of being in the binding interface.

2.2.4. Loss Function

We train the model with Binary Cross Entropy loss L_{BCE} balanced by the weight of the positive labels. In addition, we designed a regularization term, L_{Hist}, based on the intuition that the binding regions on both input point clouds should have complementary shapes [119]. To describe the shapes of the binding regions, we use D_2 shape distributions, which are the distance distributions (represented by histograms) of subsets of random points in the positive labeled point clouds. We then use the difference of distributions as a regularization loss. In particular, given the normalized distances D between all pairs of point from a random subset, we implemented a backpropagatable 10 bin histogram defined in terms of the bins‘ centers c_i and radii r_i:

$$\text{bin}(d) = \sum_{i=0}^{9} \left(\text{sigmoid}(\frac{d - c_i + r_i}{\sigma}) - \text{sigmoid}(\frac{d - c_i - r_i}{\sigma}) \right) b_i$$

$$\text{Hist}(D) = \sum_{d \in D} \text{bin}(d)$$

where σ is a coefficient (set by grid search in the training data) controlling the steepness of the sigmoid function and b_i is a 10 dimensional vector with the ith dimension set to 1 and rest to 0. Consequently, L_{Hist} is then defined as the L2 norm of the difference between D_{ligand} and D_{receptor}:

$$L_{\text{Hist}} = ||\text{Hist}(D_{\text{ligand}}) - \text{Hist}(D_{\text{receptor}})||$$

The loss function with regularization L_{Hist} is then defined by:

$$L = L_{\text{BCE}} + \alpha L_{\text{Hist}}$$

where the hyperparameter α setting the relative weight between the losses is found by grid search in the training data.

2.2.5. Training

Since the input point cloud sizes can vary significantly and the memory size of the GPU used in implementation is limited, we do not pad all input point clouds; but for particularly large protein pairs, we uniformly subsample them, without replacement, so as to reduce the cardinality of each to at most 20000. We feed one protein pair at a time into the network, performing backpropagation on the average loss of every set of complexes (16 for smaller datasets and 32 for larger ones). Thus the model is still essentially trained using mini-batch gradient descent. We use the Adam optimizer and train each model for 160 epochs with initial learning rate 0.001 decreasing by half every 20 epochs. The keep rate for the dropout layer in the segmentation block is set to be 0.7. All these parameters were determined by grid search within the training set.

2.2.6. Data Augmentation

Since there is relatively little training data available (tens to hundreds of protein complex structures in curated datasets), we sought to augment it artificially. In particular, since the spatial transformation network needs sufficient data to learn the mapping from point clouds to a canonical space, and the reference coordinate system in each structure is essentially arbitrary, a natural augmentation is to add randomly rotated instances of the training data. For each input pair of point clouds in the training set, we generated a fixed number of additional input point clouds, each randomly rotated from the original, i.e., by a random angle around a random axis. For smaller datasets, we compared the performance using 10 or 50 additional training instances for each original; for a large dataset we were only able to augment with 5 additional instances each due to memory limitations.

2.2.7. Residue-level Prediction

While PInet's predictions are in terms of surface point clouds, most methods (and indeed, subsequent experiments, e.g., site-directed mutagenesis) focus on which amino acids are in the interfaces. Thus for interpretation and direct comparison with other methods, we compute predictions for residues based on predictions for nearby surface points. In order to ensure potential representation of all atoms in a residue, we identify the closest surface points for each of its atoms and then the closest of those points over all atoms in the residue. We then average the probabilities for those nearby surface points to give the residue's probability. More precisely, for atom A_j let C_j be the corresponding k_1 closest points in the point cloud (we used $k_1 = 3$), and for residue R_k, let D_k be the k_2 closest points (we used $k_2 = 10$) from the set of points $\bigcup_{A_j \in R_k} C_j$ of atoms comprising the residue and their closest points. Then we compute the probability $P(R_k)$ that R_k is in the binding site by summing the evidence provided by the points associated with its constituent atoms.

$$P(R_k) = \frac{1}{k_2} \sum_{P_i \in D_k} P(P_i)$$

Section 2.3

Results

2.3.1. Benchmarks

We evaluated PInet on three previously collected and evaluated benchmarks, summarized in Supplementary Table 1 and Supplementary Figure 2, and associated state-of-the-art predictors. Different steps were taken by each in order to clean and reduce redundancy, as briefly summarized here.

- **DBD5**. The Protein-Protein Docking Benchmark 5.0 (DBD5) [168] is a non-

redundant set of high-resolution structures of 225 general protein-protein complexes. The redundancy is handled at the structural family level according to the Structural Classification of Proteins (SCOP) [108]. We also considered the DBD3 dataset [69], a previous version that is a subset of DBD5 of size 73. We compare our model with representative state-of-the-art methods BIPSPI [137] and PAIRpred [2], which are partner-specific predictors (i.e., for a given pair of proteins) using both sequence- and structure-based features. All models followed the same protocol of leave-one-out cross-validation. We tested on both bound conformations (i.e., the structures of the individual proteins are extracted from the complex structure), as well as on unbound conformations (i.e., the structures of the individual proteins were solved separately).

- **MaSIF** The authors of MaSIF compiled a massive dataset by combining the PRISM [107] dataset [114], the ZDock benchmark [125], PDBBind [169], and the SAbDab antibody database [41] filtered for a maximum of 30% sequence identity using psi-cd-hit [67]. In all, the original dataset used 3003 proteins for training and 359 for testing. The overall dataset was not curated for redundancy and overlap between training and testing, so for comparison we use the same test-train split, except for one small bit of curation: we discarded complexes whose binding site was smaller than 1% of the size of ligand, since from inspection these complexes have little interaction. Thus our training and testing set sizes are 2689 and 345. We compared our model with MaSIF as well as SPPIDER [127], another partner-independent predictor based on relative solvent accessibility, whose performance is included in the MaSIF publication [50]. The provided benchmark uses bound conformations.

- **EpiPred** EpiPred [84] is one of the state-of-the art antibody epitope predictors, though recently PECAN [126] outperformed it on the same dataset. The

EpiPred dataset comprises a non-redundant set of high-resolution antibody-antigen complex crystal structures, filtered from SAbDab [41] according to structural quality and dissimilarity. It includes 148 Ab-Ag complexes with Ab sequence identity less than 99% and corresponding Ag sequence identity less than 90%, and all Ags at least 50 residues. Of these complexes, 118 are used for training and 30 for testing. Here we compare our model with EpiPred along with the partner-independent epitope predictor DiscoTope [85] and the recent partner-dependent epitope predictor PECAN, which holds the current state-of-the-art performance for the EpiPred dataset, achieving 15.7% precision with 73.0% recall and AUC-PR of 65.5%. Here too, the provided benchmark uses bound conformations.

We first evaluated PInet on DBD5 in order to establish a comparison with the general state of the art in predicting how pairs of proteins interact. We also used DBD5 to show the importance of partner-specific methods, accounting for the other protein when predicting on the one protein, rather than just generally predicting what part of a protein might be recognized by any partner (see again Figure 2.1). DBD5 also provided the opportunity to assess the utility of data augmentation for our method. We then used the MaSIF dataset to evaluate our approach on this much larger and richer (though uncurated) dataset. As we can see in Supplementary Figure 2, in MaSIF, the sizes of ligands and receptors tend to be more similar compared to DBD5 and EpiPred, but the proportions of residues in interface regions (i.e., percentages of positive labels) varies substantially. We further leveraged the MaSIF dataset to explore different variations of the the PInet model, including architecture and feature choices. Finally we turned to the special case of antibody-antigen prediction and used the EpiPred benchmark to evaluate PInet's performance on this important task.

In general, due to the imbalance between positive labels (about 8%) and negative

labels (Supplementary Table 1), we advocate the use of precision and recall as performance metrics, both as single values at the 0.5 probability cut-off, as well as in a precision-recall curve. To allow comparison with some other approaches, we also compute AUC-ROC even though the classes are very imbalanced. For consistency of comparison, PInet's point-based predictions are converted to residue-level predictions as described in the methods. In order to illustrate the power of PInet in enabling further computational modeling or experimental evaluation, we provide illustrative examples of resulting segmentations.

2.3.2. DBD5

We trained and tested PInet on the DBD5 and older DBD3 benchmarks using leave-one-out cross-validation and found that, compared to representative predictors trained and tested the same way, it attains significantly better average precision, recall, and AUC-PR (Table 2.5 and supplementary table 2). For example, for the DBD5 benchmark, using a probability cutoff of 0.5, the average precision of the model trained without data augmentation is over 51% and average recall nearly 75%, both more than 10% higher than previous methods, while the average AUC-PR is almost 0.67, more than 0.20 higher. Clearly the model does a good job of identifying the binding regions, but also expands them to include some false positives. For the model trained with 10x augmentation, the performance improves slightly, while for model trained with 50x augmentation, the performance further improves substantially, attaining average precision of 54%, average recall of 82%, and average AUC-PR of 0.73. To evaluate the impact of conformational change on the predictions, we further tested PInet on the unbound structures using the 50x augmentation model. We note that BIPSI's performance was more or less the same in application to bound or unbound structures, presumably because its features are less sensitive to the conformational changes induced by binding in these test cases. As might be expected for PInet,

Table 2.1: Performance evaluation for DBD3 and DBD5 datasets, with published values for PAIRPred and BIPSI, compared to PInet with no augmentation, or augmented with 10 or 50 random rotations per training complex. Bolded entries are the best in that column for that dataset. Testing is performed on either bound (B) or unbound (U) structures, as indicated in the first column.

Struct	Method	precision	recall	AUC-ROC	AUC-PR
DBD3					
B	PInet	**0.494**	0.723	0.812	0.639
B	PInet (Aug 10)	0.480	0.732	0.846	0.669
B	PInet (Aug 50)	0.491	**0.845**	**0.867**	**0.710**
U	PAIRpred	0.371	0.419	0.774	0.341
U	BIPSPI	0.383	0.545	**0.816**	0.405
U	PInet (Aug 50)	**0.518**	**0.745**	0.775	**0.626**
DBD5					
B	BIPSPI	0.394	0.599	0.827	0.429
B	PInet	0.511	0.749	0.837	0.667
B	PInet (Aug 10)	0.523	0.755	0.851	0.685
B	PInet (Aug 50)	**0.538**	**0.824**	**0.877**	**0.734**
U	BIPSPI	0.391	0.558	**0.822**	0.410
U	PInet (Aug 50)	**0.492**	**0.723**	0.753	**0.596**

which uses features based on higher-resolution point clouds, performance did drop, but it still outperformed these other predictors in terms of both precision and recall.

Figure 2.3 illustrates epitope predictions (Supplementary Figure 3 gives paratopes) for hen egg lysozyme and antibodies illustrated in Figure 2.1. The PInet predictions clearly demonstrate a very important point: it is not just memorizing binding sites during training — while it sees the antigen and one specific antibody during training, it is still able to predict the very different interface of that antigen for different partner antibodies during testing. The result also demonstrates the importance of partner-specific prediction, as for example DiscoTope covers much of the surface as possible binding interface without distinguishing for which antibody. In contrast, the PInet predictions leverage antibody specific models to distinguish different binding regions and support prioritization and follow-up studies for the different antibodies.

Figure 2.4 visualizes the PInet interface region predictions for the median per-

formance complexes from the DBD5 Enzyme-Partner category test cases. Here too — even at the median level of performance — PInet is able to provide high quality interface region segmentations. Overall (Supplementary Figure 4), the prediction performance is consistent across different types of protein interactions, suggesting that PInet is learning robust, general models of interaction specificity.

2.3.3. MaSIF

The MaSIF benchmark is substantially larger, and provides more diverse but uncurated training and testing data. The MaSIF paper also presents results for a model using geometric features only, as compared to the one using both geometry and physicochemical features. We trained PInet following the same protocol and, as the top rows in Table 2.2 indicate, PInet outperforms MaSIF when using just geometric features, and attains the same performance when using both geometric and physicochemical properties. The relatively better performance of PInet on geometry alone leads us to conjecture that PInet could benefit from a different encoding of physicochemical features or a more direct aggregation of the features (e.g., as convolution would do), In addition to the full dataset, the MaSIF paper also includes a comparison to SPPIDER [127] on predicting for the subset of single-chain transient interactions. We also tested on the transient interactions subset; the bottom section of Table 2.2 shows that we again match MaSIF's performance in terms of AUC-ROC. For completeness, since the dataset is quite imbalanced, we also provide AUC-PR for PInet; though a comparison can't be made to the other methods, we do see that it is substantially lower than we obtained for the DBD5 dataset, suggesting that a carefully curated training and testing regime can yield better performance.

Finally, we used the MaSIF dataset to compare different versions of the PInet modeling approach; performance is summarized in Supplementary Table 3. The top two rows shows the full model and the geometry-only model, as already presented

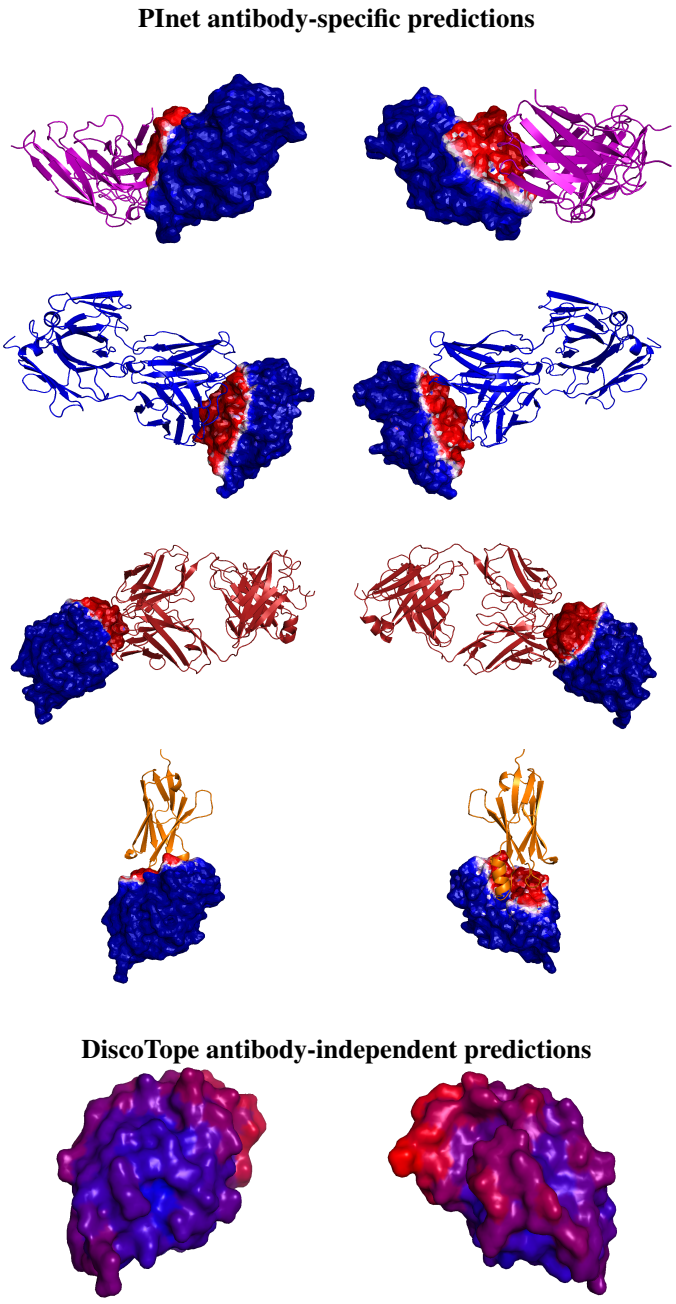

Figure 2.3: Binding interface region prediction for one protein (hen egg lysozyme) with four different antibodies. (Top four rows) PInet predictions, from two different viewpoints. The heatmap shows predicted probability of being in the interface, with darker red for higher probability and darker blue for lower. (Bottom row) Discotope prediction for 3LZT, again with a heatmap showing predicted probability (darker red higher, darker blue lower).

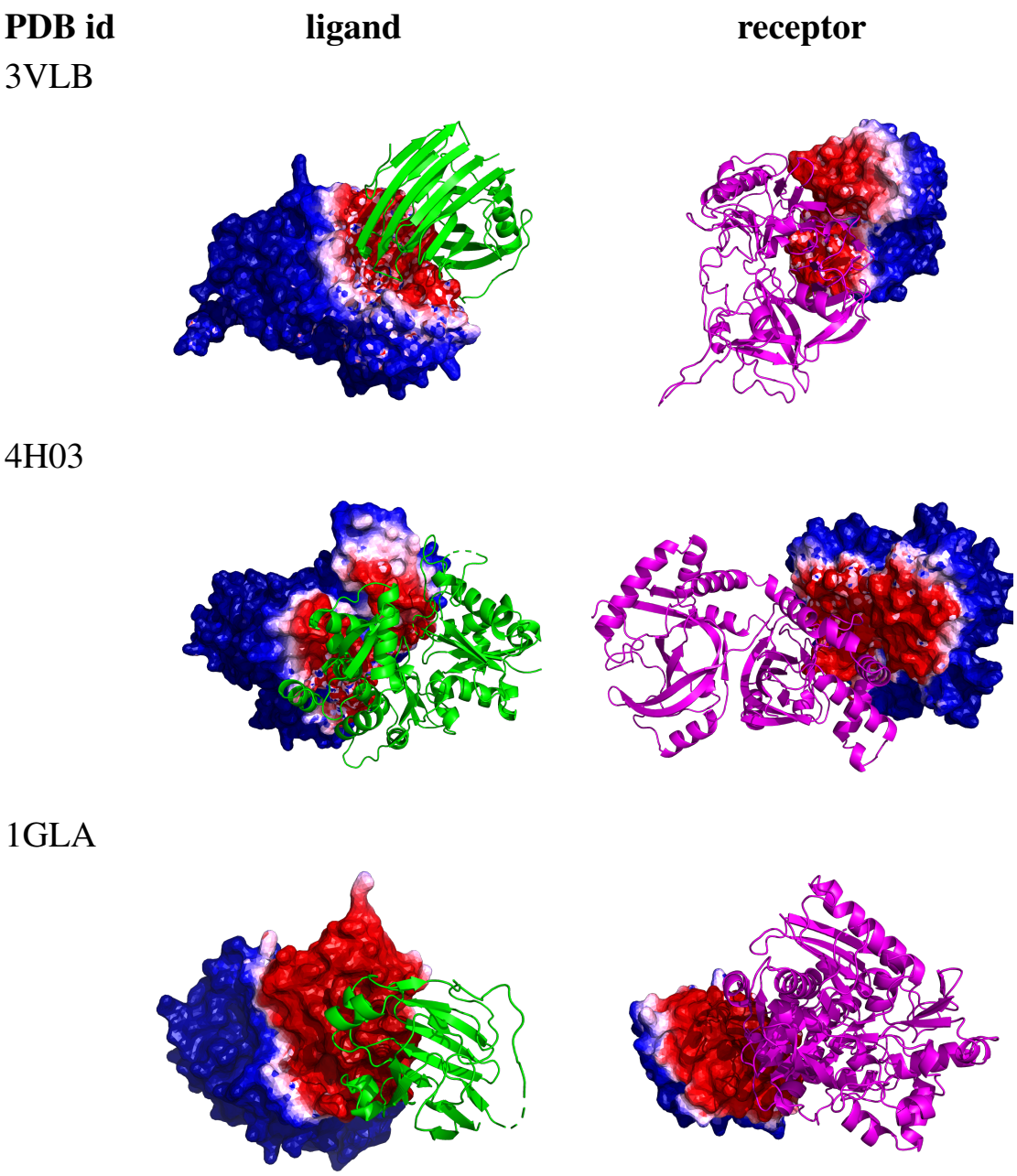

Figure 2.4: Segmentation visualization for three example (median PInet prediction performance) Enzyme-Partner pairs from DBD5: Enzyme-Inhibitor 3VLB (top), Enzyme-Substrate 4H03 (middle), and Enzyme complex with a regulatory or accessory chain 1GLA (bottom). Heatmaps for the partner indicate the predicted probability of being in the interface, with darker red for higher probability and darker blue for lower.

Table 2.2: Performance (area under the PR and ROC curves) on the Masif benchmark. (top) full test set, with each model using geometric features and physicochemical features. (middle) full test set, with each model using geometric features only. (bottom) transient interaction subset.

Test case	Method	AUC-PR	AUC-ROC
Full	MaSIF	n/a	0.87
Full	PInet	0.45	**0.88**
Full	MaSIF (geom only)	n/a	0.68
Full	PInet (geom only)	0.30	**0.75**
Transient	SPPIDER	n/a	0.65
Transient	MaSIF	n/a	0.81
Transient	PInet	0.30	**0.82**

in Table 2.2. The next two rows summarize the impact of adding each of our two physicochemical features separately; each helps, though hydrophobocity helps more on this dataset. To evaluate the effects of point cloud resolution, we also evaluated a model using at most 2000 points, instead of the default 20,0000 points. As shown in Supplementary Figure 2, most point clouds are in the 10,000 to 20,000 range, so this subsampling reduces the grid resolution and consequently hurts performance all around. Finally, we compared the difference between models with or without the data augmentation. The extent of augmentation we could perform was limited due to memory constraints, and the dataset was already quite large, so augmentation had little effect on the performance here.

2.3.4. EpiPred

The final dataset covers the important special case of antibody-antigen interactions. While our model simultaneously predicts both interfaces, the epitope (the part of the antigen recognized by the antibody) is usually most important, so we focus on comparison with the state-of-the-art antibody-specific epitope predictors PECAN [126] and EpiPred [84], along with the state-of-the-art antibody-independent DiscoTope [85], using Epipred's training and testing sets.

Table 2.3: Epitope prediction performance on the EpiPred test set.

Method	Precision	Recall	AUC-ROC	AUC-PR
EpiPred	0.136	0.436	NA	NA
DiscoTope	0.214	0.110	NA	NA
Sppider	0.153	0.363	NA	NA
PECAN	0.157	0.730	0.655	0.226
PInet [EpiPred]	0.181	**0.931**	0.654	0.291
PInet [EpiPred & Aug]	**0.216**	0.774	**0.687**	**0.368**
PInet [DBD5 Aug] ave	0.206	0.752	0.651	0.321

Table 2.3 summarizes the performance of the various approaches. The basic PInet approach, training and testing on the EpiPred dataset alone, is directly comparable to the state-of-the-art PECAN, and we see that it substantially outperforms it (as well as the other methods) in terms of recall while also doing better on precision. Data augmentation further improves precision and also achieves the best AUC-PR and AUC-ROC. Strikingly, the PInet models trained above on augmented DBD5 data (performance numbers averaged over all models trained during the leave-one-out cross-validation) achieve performance comparable to that of the other state of the art methods, though suffer a loss in recall compared to the model trained directly on EpiPred data. This result implies that our model robustly learns generalizable determinants of protein-protein binding interfaces.

Figure 2.5 visualizes the PInet epitope predictions for representative good and not-so-good test cases. Corresponding paratope predictions are included in Supplementary Figure 5. Consistent with the performance metrics, PInet identifies most or all of the epitope regions, but extends the segmentation out a bit further beyond them. By the median performance example, we see that PInet also produces a secondary binding site; these false positives would need to be ruled out by subsequent computational or experimental analysis. Finally, we evaluated the overall performance for simultaneous paratope and epitope prediction (Supplementary Table 4). We see that

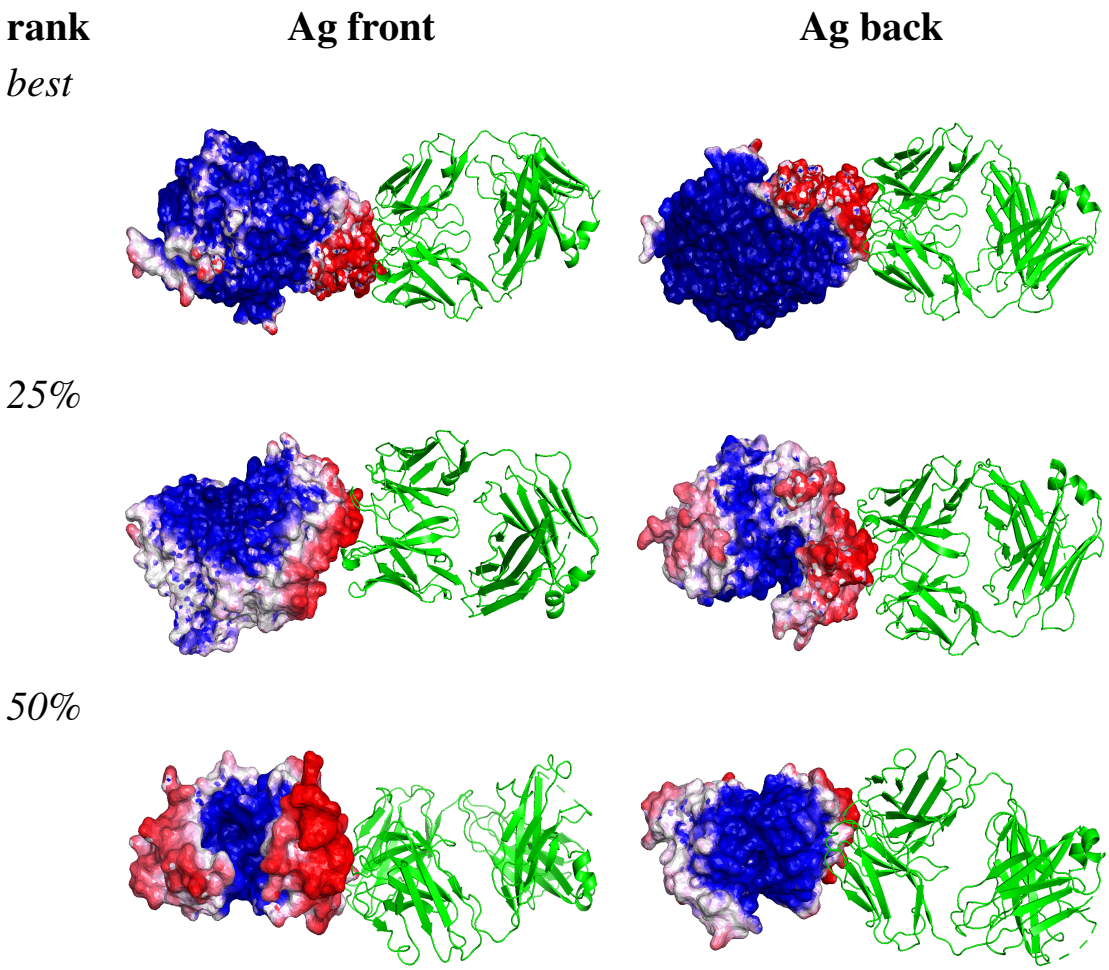

Figure 2.5: Segmentation visualization for 3 example Ab-Ag pairs: best 4JR9, top 25% 3LIZ, and 50% 3RAJ (according to AUC-PR). For these three pairs, precision ranges from 13% to 23% and recall ranges from 65% to 93%. Green structures are antibodies and gray structures are antigens, while heatmaps for the partner indicate the predicted probability of being in the interface, with darker red for higher probability and darker blue for lower.

indeed the precision and recall for combined epitope+paratope prediction tends to be higher than that for epitope prediction since, as discussed, paratope prediction is generally easier than epitope prediction, due to the structural and functional similarity of the complementarity determining regions of antibodies.

Section 2.4

Discussion

In order to improve the state of the art in predicting how proteins interact, aiming to scale up to support large-scale analyses, we bridge physical modeling approaches and data-driven approaches, training a data-driven model using the same types of features leveraged by physically-based modeling methods. Our unified geometric deep neural network encodes geometry and physicochemical properties by way of point clouds and leverages this encoding to learn how to recognize the complementarity underlying molecular recognition. This leads to interface region prediction that outperforms state of the art methods in precision, recall, and AUC-PR. Moreover, our model could be directly used as a protein surface fingerprint for more applications such as pose recovery.

The model trained on the general DBD5 dataset performed quite well on the specific EpiPred dataset, but could potentially benefit from additional training with antibody-antigen data. Thus we evaluated the utility of transfer learning, pre-training a model with DBD5 and then fine tuning it with EpiPred training data. The transfer learning process did boost the performance on the EpiPred test set relative to a model trained on EpiPred alone with no augmentation, but the benefit was not significant compared to that of a model trained on EpiPred with augmentation (Supplementary Table 4). This suggests that, at least as far as this dataset goes, the model learned from general protein complexes already represents antibody-antigen recognition as

well as it can.

While it can be very challenging to discern what a neural network has learned and how exactly it is representing information, we sought to explore some of the key properties of PInet. Since the global vector is summarized by a pooling layer, we plotted those points that most strongly activates channels in the global vector for a couple of example proteins, and found that PInet's global feature extraction layer is essentially sampling landmarks on the protein surface and using them to summarize the protein (Supplementary Figure 6a). The following MLP layer then segments the surface based on predictions from these points as to where interactions are most likely. We also noticed that the global vectors of different classes of interaction in DBD5, Ab-Ag vs. Enzyme-Partner, are distinguishable when projected into 2D via t-SNE embedding (Supplementary Figure 6b). Even though the model was not trained for this type of classification, it appears to be representing differently these different types of interfaces. These preliminary investigations point to the potential value of unsupervised pre-training on single proteins (for which there is much more data) and using a better encoding as the basis for learning models of interaction specificity.

The strength of PInet lies in its point cloud representation of protein surfaces and its ability to learn representations of complementarity in interacting surfaces. Table 2.2 and Supplementary Table 3 show that PInet receives much less benefit from augmenting its geometric representation with biophysical properties, compared to MASIF. This could be due to the discrete encoding of electrostatic features employed by PInet, which lacks a convolutional kernel that would more naturally capture electrostatic locality. Future work may seek to incorporate a CNN like approach into the point cloud architecture, or use a different encoding of electrostatics to better leverage this important information as MASIF does.

The point cloud representation makes PInet more sensitive to conformational

change upon binding (B vs. U in Table 2.5). In general, physically-based methods such as docking depend on such details, while the features used by and representations learned by data-driven methods may not (e.g., solvent accessibility features may be generally unaffected by modest conformational change). As a hybrid physics/data-driven approach, PInet falls somewhere in the middle. Thus an interesting angle for future work is to train models that learn the features of complementary surfaces while robustly accounting for deformations to those surfaces induced during binding.

PInet pursues a partner-specific prediction approach, reasoning that in theory, one partner's sequence "encodes" information regarding what it will recognize on the other, and nature is able to "decode" this information reliably. Thus it is worthwhile to attempt to learn such a representation, which could be useful, e.g., to identify which of a set of isolated antibodies target which sites on the antigen, which representative subset is thus worth functionally characterizing, which antibodies are more likely to be neutralizing, and so forth. The partner-specific approach clearly pays off in the hen egg lysozyme example, where PInet deconvolves the different antibody specificities and gives much better localizations of their putative epitopes compared to the partner-independent method. However, the overall precision across the benchmarks still remains low, and there is clearly still much work to be done to achieve sufficiently accurate predictions. While partner-specific methods have more information than partner-independent ones, simply knowing the partner doesn't necessarily help all by itself, since it isn't known *a priori* which part of the partner's surface is the interface. In fact, simultaneous prediction of both interfaces goes a long way toward solving the docking problem. Indeed, we hope that further development of data-driven methods explicitly encoding geometric and biophysical complementarity may indeed enable direct prediction of binding modes with high precision and recall over large diverse sets of protein partners.

Section 2.5

Acknowledgements

We would like to thank Srivamshi Pittala for helpful comments on this work. We would also like to thank Dartmouth Research Computing for help with the Discovery cluster.

Section 2.6

Funding

This work was supported in part by NIH grant 2R01GM098977.

Section 2.7

Supplementary

Table 2.4: Benchmark dataset sizes and binding interface fractions. $\mu(L_1)$ and $\sigma(L_1)$ are respectively the mean and standard deviation of fraction of positive labels (i.e., interface region).

Dataset	Training size	Testing size	$\mu(L_1)$	$\sigma(L_1)$
DBD5	189	36	0.083	0.034
DBD3	60	13	0.088	0.026
MaSIF	2689	345	0.125	0.060
EpiPred	118	30	0.056	0.018

Table 2.5: Additional performance evaluation for DBD3 and DBD5 datasets. Bolded entries are the best in that column for that dataset. PInet significantly outperformed the others in precision, recall, and AUC-PR, the most important metrics for interface prediction.

	precision	recall	specificity	NPV	AUC-ROC	AUC-PR	MCC
DBD3							
PAIRpred	0.371	0.419	**0.898**	0.925	0.774	0.341	0.311
BIPSPI	0.383	0.545	0.887	**0.938**	0.816	0.405	0.373
PInet	**0.494**	0.723	0.762	0.894	0.812	0.639	0.428
PInet (Aug 10)	0.480	0.732	0.805	0.926	0.846	0.669	0.459
PInet (Aug 50)	0.491	**0.845**	0.751	0.938	**0.867**	**0.710**	**0.497**
DBD5							
BIPSPI	0.391	0.558	**0.889**	**0.940**	0.822	0.410	0.385
PInet	0.511	0.749	0.778	0.900	0.837	0.667	0.459
PInet (Aug 10)	0.523	0.755	0.784	0.911	0.851	0.685	0.496
PInet (Aug 50)	**0.538**	**0.824**	0.793	0.925	**0.877**	**0.734**	**0.526**

Table 2.6: Performance of different variants of PInet on the full MaSIF dataset, evaluated by precision, recall, AUC-ROC, and AUC-PR. The full model is compared to models using geometry alone or geometry plus either of the two physicochemical properties, a model with the point cloud downsampled to 2000 points, and a model incorporating the regularization term.

Method	Precision	Recall	AUC-ROC	AUC-PR
full	0.31	0.63	0.88	0.45
aug 5	0.34	0.61	0.89	0.45
geometry only	0.21	0.58	0.75	0.30
geometry + electrostatics	0.23	0.58	0.78	0.32
geometry + hydrophobicity	0.24	0.62	0.82	0.37
subsample 2000	0.28	0.53	0.81	0.38

Table 2.7: Performance on simultaneous epitope+paratope prediction

Method	Precision	Recall	AUC-ROC	AUC-PR
PInet [EpiPred alone]	0.193	**0.898**	0.751	0.330
PInet [EpiPred alone Aug]	**0.261**	0.815	**0.795**	**0.447**
PInet [DBD5 & tuned]	0.250	0.864	0.778	0.412

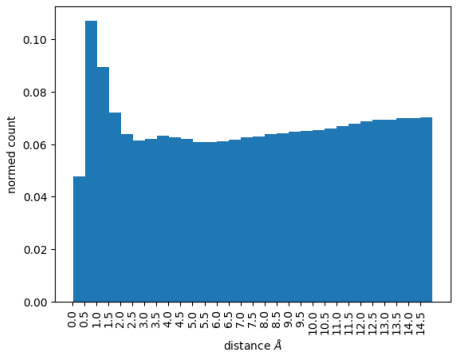

Figure 2.6: Distribution of pairwise shortest distances from each point on one protein to the closest on the other.

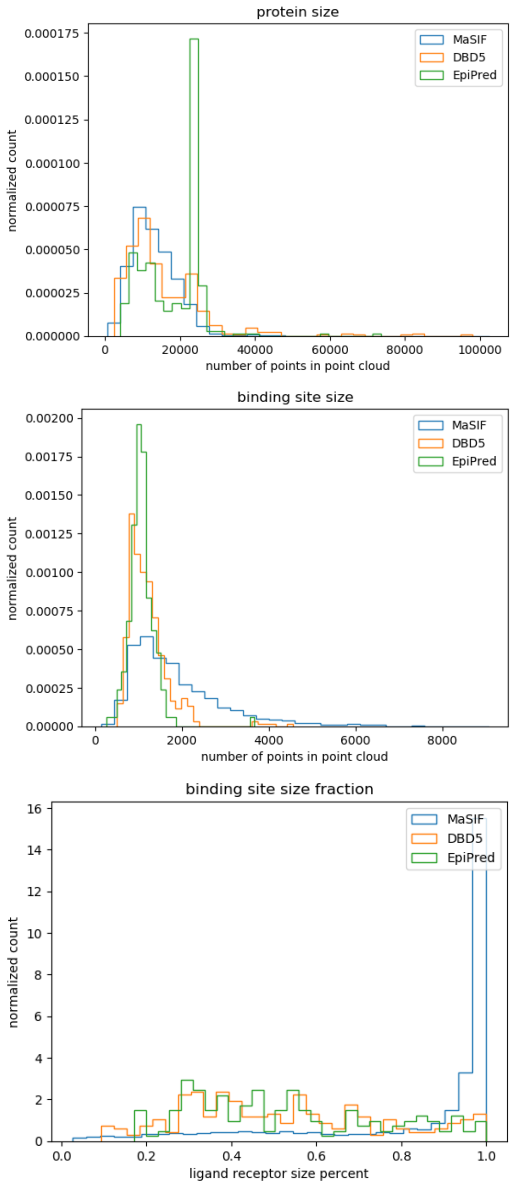

Figure 2.7: Characterization of datasets in terms of sizes of proteins and their binding regions. Counts are normalized so that the bin areas sum to one. Protein size and binding site size are characterized by number of points in the point cloud, and the relative size of the binding interface is computed by number of points in interface over total number of points in protein.

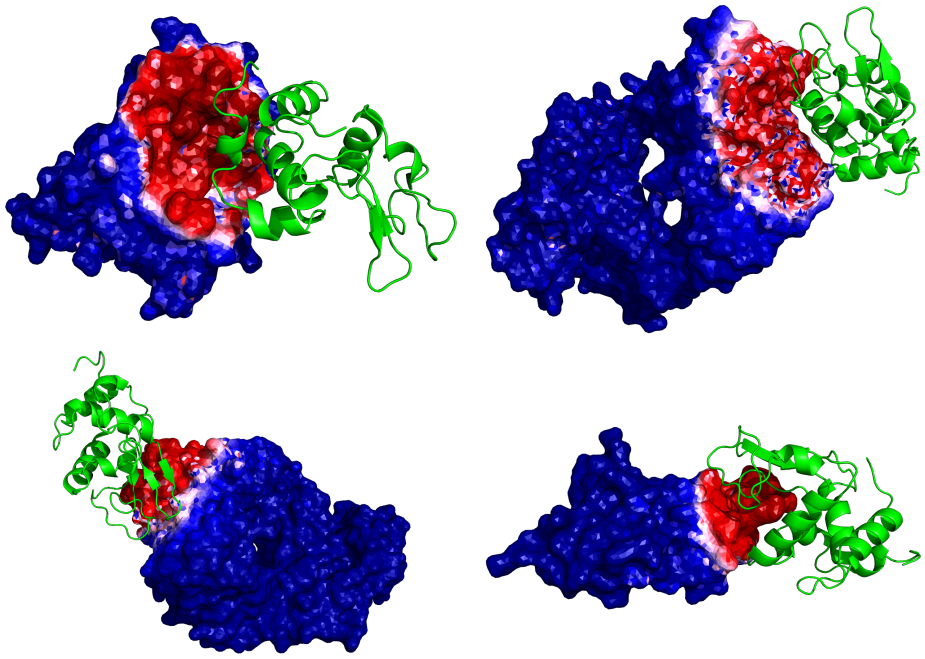

Figure 2.8: Segmentation visualization for paratope prediction of three anti-hen egg lysozyme antibodies (left to right, top to bottom: PDB ids IBVK, 1DQJ, 1MLC, and 2I25). Hen egg lyzosyme (PDB id 3LZT is rendered as cartoon, while antibodies are rendered as surface heatmaps according to the predicted probability of being in the paratope, with darker red for higher probability and darker blue for lower.

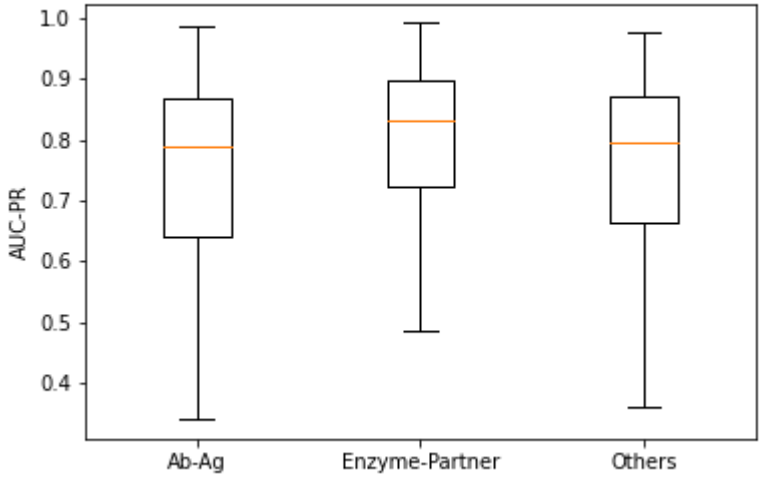

Figure 2.9: Boxplot of AUC-PR by DBD5 protein interaction class.

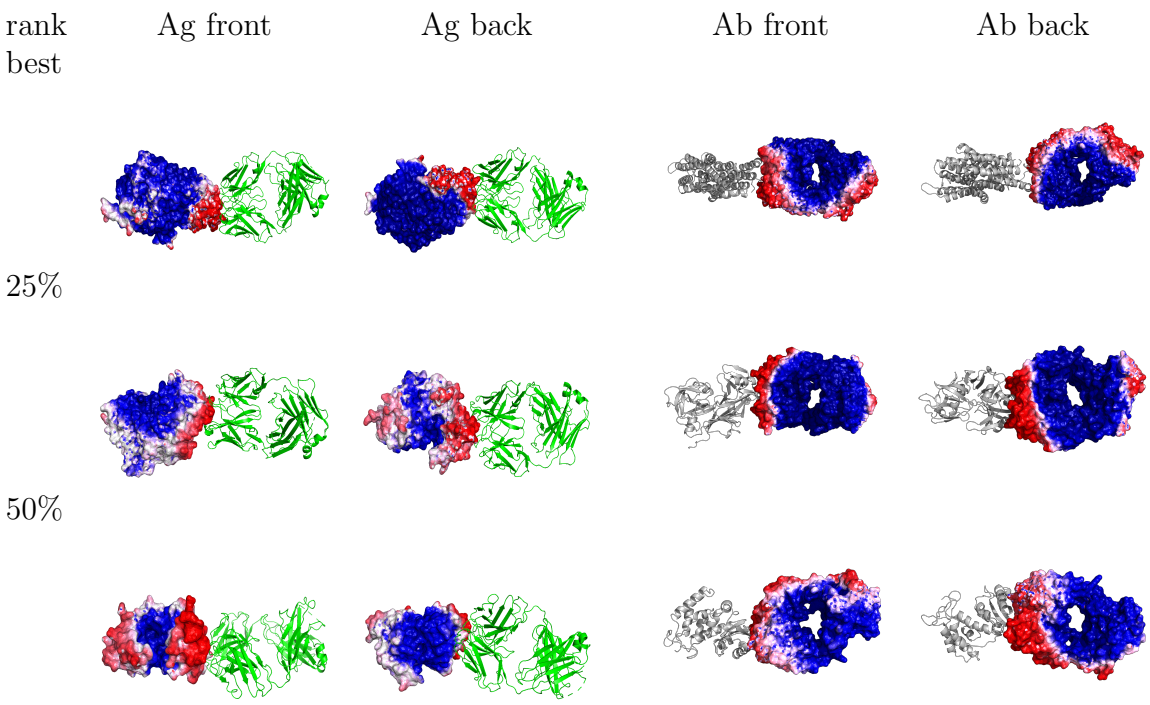

Figure 2.10: Segmentation visualization for three example Ab-Ag pairs: best 4JR9, top 25% 3LIZ, and 50% 3RAJ (according to AUC-PR). For these three pairs, precision ranges from 13% to 23% and recall ranges from 65% to 93%. Green structures are antibodies and gray structures are antigens, while heatmaps for the partner indicate the predicted probability of being in the interface, with darker red for higher probability and darker blue for lower.

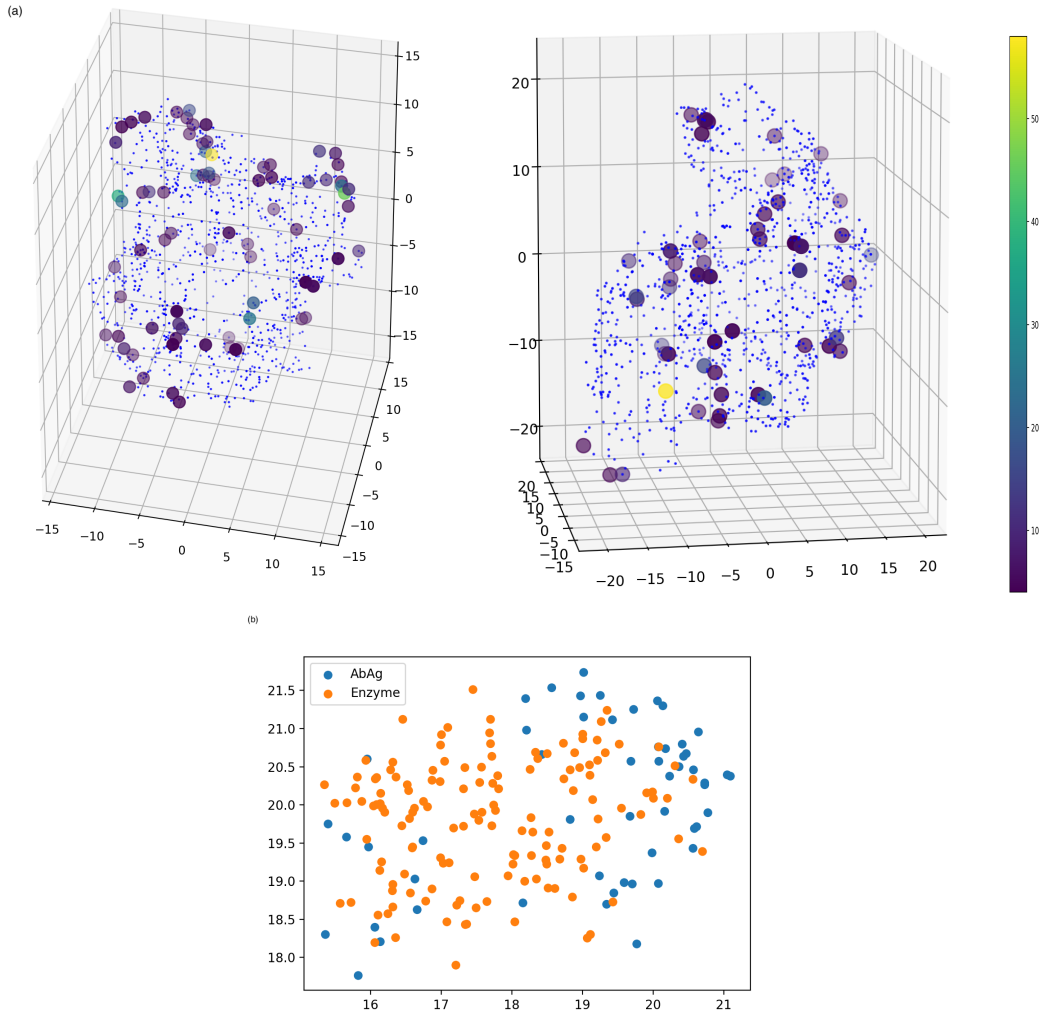

Figure 2.11: (a) Surface points (blue) for two example proteins (left: PDB id 1DKS; right: PDB id 1HCL), highlighting the points with the highest activation in the global vectors (colorscale: number of channels activated). (b) t-SNE plot of global vectors for proteins from the DBD5 dataset colored by the type of interaction (blue: Ab-Ag; orange: Enzyme-partner).

Chapter 3

Attention please: modeling global and local context in glycan structure-function relationships

Glycans are found across the tree of life with remarkable structural diversity enabling critical contributions to diverse biological processes, ranging from facilitating host-pathogen interactions to regulating mitosis & DNA damage repair. While functional motifs within glycan structures are largely responsible for mediating interactions, the *contexts* in which the motifs are presented can drastically impact these interactions and their downstream effects. Here, we demonstrate the first deep learning method to represent both local and global context in the study of glycan structure-function relationships. Our method, glyBERT, encodes glycans with a branched biochemical language and employs an attention-based deep language model to learn biologically relevant glycan representations focused on the most important components within their global structures. Applying glyBERT to a variety of prediction tasks confirms the value of capturing rich context-dependent patterns in this attention-based model: the same monosaccharides and glycan motifs are represented

differently in different contexts and thereby enable improved predictive performance relative to the previous state-of-the-art approaches. Furthermore, glyBERT supports generative exploration of context-dependent glycan structure-function space, moving from one glycan to "nearby" glycans so as to maintain or alter predicted functional properties. In a case study application to altering glycan immunogenicity, this generative process reveals the learned contextual determinants of immunogenicity while yielding both known and novel, realistic glycan structures with altered predicted immunogenicity. In summary, modeling the context dependence of glycan motifs is critical for investigating overall glycan functionality and can enable further exploration of glycan structure-function space to inform new hypotheses and synthetic efforts. This paper is under final editing and is available on Biorxiv `https://www.biorxiv.org/content/10.1101/2021.10.15.464532v1.full.pdf`.

Section 3.1

Introduction

Glycans are complex oligosaccharides often presented in branched structures attached to proteins, lipids, and RNA, with critical roles in a diverse range of biological processes [162, 48]. Glycans mediate these processes through two general mechanisms [161]: (1) specific interaction and recognition of glycoforms by glycan binding proteins, such as the targeting of sialoglycans on the surface of animal cells by influenza hemagglutinin to facilitate viral entry [36], and (2) structural and biophysical effects such as the critical modulation of antibodies' downstream effector functions by N-linked glycans [72]. Subtle modifications in glycan structures independent from functional epitopes can drastically alter glycan function, as seen in the large structural and functional shifts introduced by core fucosylation [113, 140]. Additionally, lectin-glycan interactions have recently been shown to be dependent on the surround-

ing global structural context in which the appropriate binding epitopes are presented [91, 53], e.g., as demonstrated by the glycan binding preferences of *Maackia amurensis* lectin I (MAL-I) (Figure 3.1, left). As glycan synthesis, purification, and presentation techniques advance and more detailed interrogations of glycans' structure-function relationships become available [91, 53], methods are needed to analyze and explore (Figure 3.1, middle and right) complex context-dependent aspects of glycan structure-function relationships.

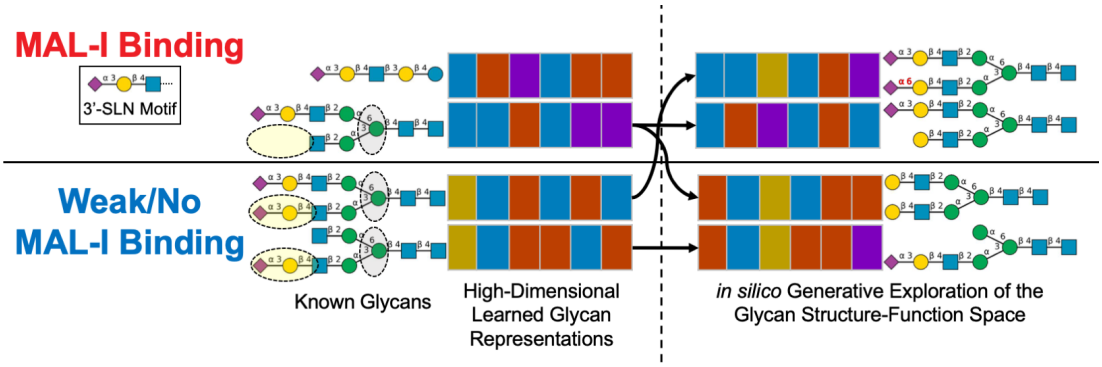

Figure 3.1: **Schematic of an analytical framework to investigate context-dependent structure-function relationships in glycans. (left)** Local context (functional glycan motifs such as 3' sialyllactosamine (3'-SLN)) and global context (non-adjacent features indicated by dashed circles) simultaneously play critical roles in the functionalities of glycans, as demonstrated here by the glycan-binding preferences of *Maackia amurensis* lectin I (MAL-I), which recognizes 3'-SLN on linear glycans or on the α1,6 branch of biantennary glycans (grey circles), but only when the opposing α1,3 branch does not present the 3'-SLN motif (yellow circles). **(middle)** As schematically illustrated, our glyBERT approach captures such contextual relationships with high-dimensional glycan representations that can predict properties of new glycans. **(right)** GlyBERT also enables generative exploration of glycan structure-function space to provide insights into learned patterns and probe glycan diversity beyond the practical reach of synthesis, thereby generating novel hypotheses and informing experimental design.

Analyses of experimental measures of glycan functions offer a means to efficiently capture the components of diversity most relevant to functionality while also generalizing to other glycans, enabling the prediction of properties of previously untested glycans or even the generation of entirely novel glycans (Figure 3.1), thereby com-

plementing and extending the impact of synthetic capabilities which remain limited relative to the scale of glycan diversity [76, 90]. To date, most analytical approaches have focused on identifying common structural motifs [31] among glycans with shared properties, e.g., via subtree mining [61, 29], motif enrichment analysis [81], or multiple alignment approaches [65]. These methods have been very beneficial in formalizing high level glycan-binding preferences of proteins [60]. They will continue to retain great utility due to their high interpretability, but this interpretability comes at the cost of reduced ability to capture more complex and context-dependent patterns. In general, deep learning approaches are attractive for building detailed and powerful representations from complex information [174]. The inceptive application of deep learning methodology directly to glycans, SweetTalk [12], used a "glycoword"-based language model to successfully demonstrate that deep modeling could capture evolutionary and functional patterns in glycan structures [14]. The follow-up SweetNet approach [20] obtained increased prediction performance by utilizing graph-based glycan encodings and convolutional neural networks, thereby addressing SweetTalk's failure to capture the branched nature of glycan structures. While convolutional approaches can capture local context, they do not preserve global relationships [164] which are critical in glycan functionality [91, 53]. Another very recent effort, GlyNet [22], implemented a fingerprinting approach to predict relative strengths of protein-glycan interactions, but faces similar limitations as SweetTalk & SweetNet in terms of disregarding global structural context.

In order to capture local and global context-dependent patterns driving glycan structure-function relationships, we introduce here a new method based on the powerful deep language model known as BERT (Bidirectional Encoder Representations from Transformers) [39]. Our implementation, called glyBERT, utilizes a flexible glycan encoding based on a branched biochemical language, and presents encoded gly-

cans to an attention-based transformer network architecture [164], thereby learning to pay attention to particular local patterns that are important within particular global contexts. GlyBERT effectively learns glycan representations that capture functional pieces of glycan structural diversity in their local contexts while also incorporating the critically important global contexts of entire structures. This enables the model to obtain state-of-the-art performance in predicting properties of other glycans. Furthermore, glyBERT supports a novel proof-of-principle generative process to explore glycan structure-function space, providing insights into what the model has learned as determinants of different glycan properties, as well as potentially informing experimental design and focusing *in vitro* synthesis efforts [95].

Section 3.2

Results

GlyBERT was trained on 80% of a set of 16,048 glycans from the SugarBase database [13], labeled with glycoprotein linkage, immunogenicity, and taxonomic origin. The glycans from the withheld 20% were used to explore the context-dependent representations learned by glyBERT relevant to these properties (subsection 3.2.1) and to quantitatively evaluate prediction performance (subsection 3.2.2). GlyBERT was further trained and evaluated for prediction of lectin-glycan recognition using glycan microarray data for 20 separate lectins (subsection 3.2.2). Finally, glyBERT was leveraged to generatively explore glycan structure-immunogenicity space in a case study application of our new algorithm for generative optimization of glycan structure-function relationships (subsection 3.2.3).

3.2.1. Attention-based modeling with glyBERT learns representations capturing effects of local and global structural context

GlyBERT, like all machine learning methods, essentially transforms its training data into a novel representation (schematically illustrated in Figure 3.1), enabling it to generalize to new instances (here glycans) based on how their representations compare to the training examples. In order to make high-quality predictions about properties of the new glycans, it is critical that the learned representations suitably capture important, generalized components of glycan structures. We explored the extent to which the glycan representations that glyBERT learned from the training set generalized to new glycans in the held-out set, first in terms of global structural contexts of whole glycans, and then in structural contexts of constituent monosaccharides. Since the learned representations within the deep learning model are hard to interpret on their own, they were transformed and visualized via Uniform Manifold Approximation and Projection (UMAP) dimensionality reduction [100].

GlyBERT learned glycan representations that effectively separated out the effects of local structural motifs based on global structural contexts, differentiating glycan structures on the basis of immunogenicity, glycoprotein linkage, and taxonomic origin (Figure 3.2). The immunogenicity of the same immunogenic motifs was modulated by global structural context captured in the glyBERT-learned representations: Lewis-Y antigen-presenting immunogenic glycan SBID3980 was grouped closer to other immunogenic glycans while the exact same structure with the addition of a reducing-end terminal β-linked glucose (non-immunogenic glycan SBID4613) was grouped separately with other non-immunogenic glycans. Similarly, the immunogenic I antigen (SBID5978) is no longer immunogenic when presented on an O-GalNAc glycan where either the reducing-end terminal GlcNAc is substituted for a GalNAc (SBID2468) or a GalNAc residue is added to the reducing end (SBID12833), with clear separation

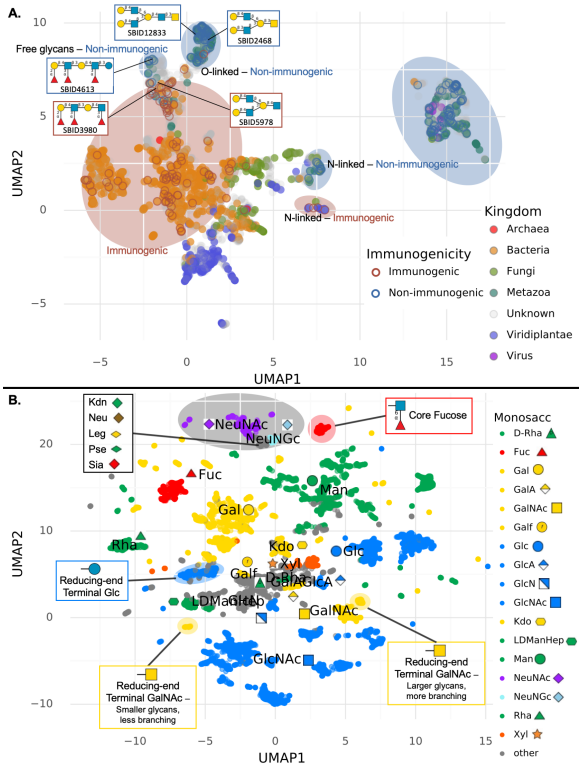

Figure 3.2: **GlyBERT-learned glycan representations capture global and local contextual relationships that are consistently manifested in an independent validation set of glycan structures. (A)** This UMAP visualization places withheld glycans with similar glyBERT representations close to each other, colored by taxonomic origin at the kingdom rank and outlined by immunogenicity. Clusters of glycans are labeled and circled, showing that glyBERT grouped the previously unseen glycans by shared immunogenicity and/or glycoprotein linkage. For further interpretation, some example glycan structures from both training and testing tests are provided, illustrating that glyBERT's representations for otherwise very similar glycans were shifted based on the influence of global context on immunogenicity (indicated by colored boxes); e.g., non-immunogenic SBID4613 and immunogenic SBID3980 both display the Lewis-Y antigen motif but in different contexts, and immunogenic SBID5978 and non-immunogenic SBID2468 & SBID12833 display the I antigen motif in different contexts. **(B)** This UMAP visualization places individual constituent monosaccharides of withheld glycans close to other monosaccharides with similar glyBERT representations, with the most common ones colored and labeled. It is apparent that the same monosaccharides are represented similarly unless their local or global structural contexts differ. GlyBERT represented fucose in a core fucosylation local context (circled in red) differently from other fucose residues, as it did for glucose in a reducing-end terminal local context (circled in blue). O-linked reducing-end terminal GalNAc residues formed two distinct clusters (circled in yellow), separated by the size and degree of branching of the corresponding glycans (global context). Additionally, glyBERT grouped together the distinct but chemically similar nonulosonic acid monosaccharides (grey circle) based on similar contextual presentations.

between these glycan representations matching their immunogenic states based on the global context of the I antigen presentation, despite high similarity between the structures.

In addition to evaluating similarities of entire glycans, constituent monosaccharides can be compared based on their individual "contributions" to the overall glycan representations. The constituent monosaccharide representations still capture their local and global context due to the attention-based architecture of glyBERT. We see that the same monosaccharide components of the withheld glycans generally shared similar representations (Figure 3.2), although glyBERT representations also captured dramatic differences in the local glycan context in which they appeared. This can be seen in the distinct groupings of fucose saccharide residues in a core fucosylation context from the rest of the fucose residues, as well as the separation of glucose as a reducing-end terminal saccharide from the rest of the glucose residues. These local contexts are clearly represented as separate from other occurrences of the same monosaccharides, as would be expected from a good representation, since reducing-end terminal glucose is sufficient to mitigate the immunogenicity of a Lewis-Y antigen-presenting glycan (Figure 3.2) and core fucosylation can induce significant conformational and functional changes in glycans [113, 140]. Learned monosaccharide representations also captured differences in the global context of their corresponding glycan structures, seen in the two separate groups of reducing-end terminal O-linked GalNAc residues distinct from other GalNAc residues based on distinct local contexts and separated from each other based on global structural contexts depending on the size and degree of branching of the entire glycan. On the flip side, different and separate monosaccharides that appear in very similar global and local contexts had similar representations as learned by glyBERT, indicating generalizabilty to novel glycan structures. This can be seen in the grouping composed exclusively of nonu-

losonic acid monosaccharides, the larger class containing sialic acid monosaccharides and other negatively charged 9-carbon monosaccharides [163]. Notably, representations of the related 8-carbon Kdo monosaccharide grouped separately from the nonulosonic acid monosaccharides despite shared functional groups and biosynthetic pathways [163].

3.2.2. Contextual glycan representations learned by glyBERT enable state-of-the-art predictive performance

As a demonstration of the utility of these learned context-dependent glycan representations, glyBERT's ability to predict glycoprotein linkage state, immunogenicity, and taxonomic origin for the glycans from the withheld 20% from SugarBase was compared to that of three recent deep learning-based approaches: SweetTalk [12, 14], SweetNet [20], and GlyNet [22]. All models predicting glycoprotein linkage performed very well, with glyBERT and SweetTalk achieving 99% accuracy, suggesting that this structural association is fairly easily represented. However, for immunogenicity, glyBERT (98.0% accuracy) substantially improved on SweetNet (94.6%) and GlyNet (95.4%), which had improved on the first-generation SweetTalk (91.7%), indicating benefit from explicitly accounting for branched structures (as done by SweetNet, GlyNet, and glyBERT) and further benefit provided by the use of attention-based modeling to capture context (glyBERT). For taxonomic origin classification (Figure 3.3), glyBERT displayed the highest accuracy at each rank, with greater increases in performance compared to SweetNet at the phylum, class, and order taxonomic ranks.

To further evaluate the benefit of attention-based glycan modeling for complex structure-function relationships, the prediction of functionally significant lectin-glycan interactions was considered, since recent studies have shown the glycan-binding preferences of lectins to be more context-dependent than previously understood [91, 53]. GlyBERT was compared to GlyNet [22] and a recent, well-performing motif-centric

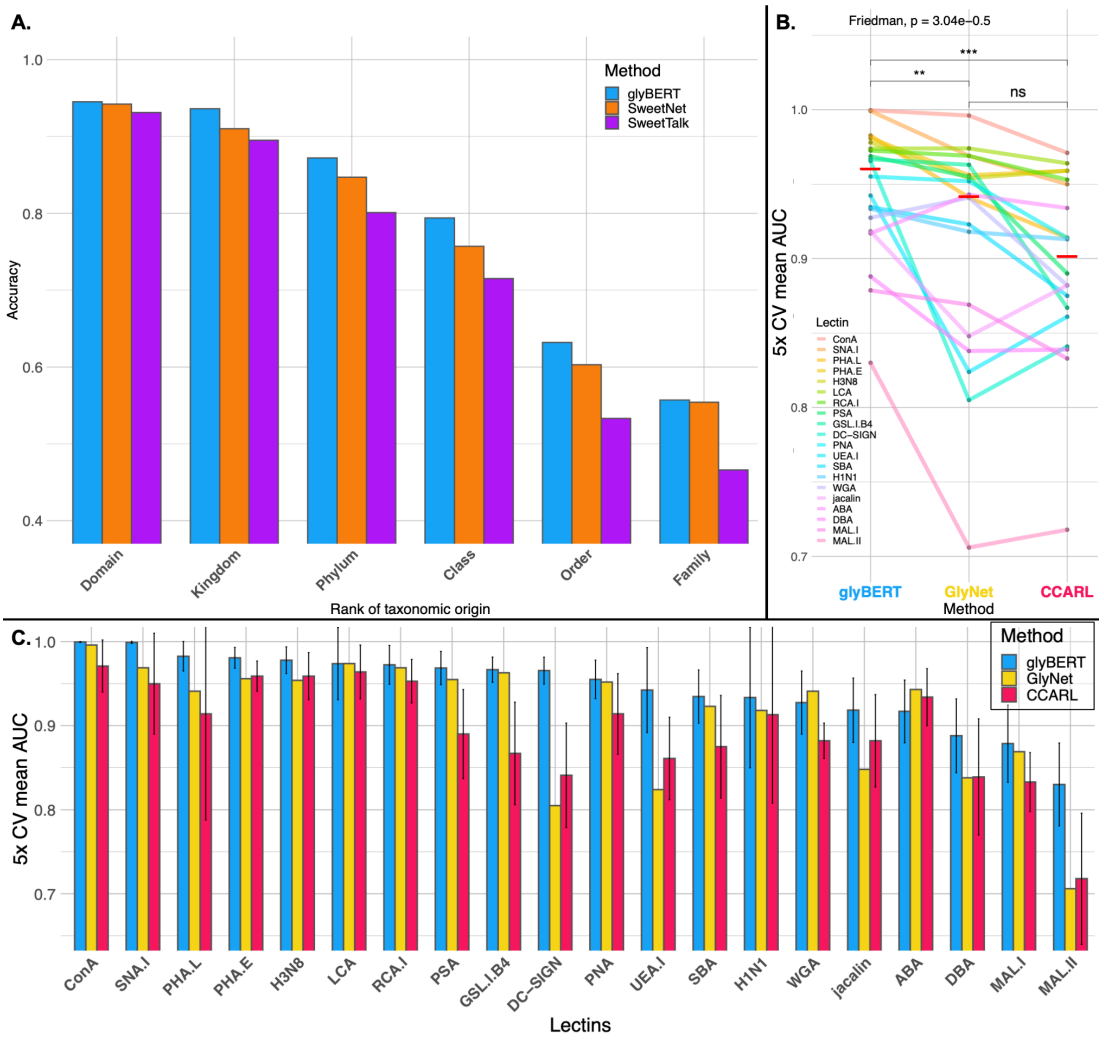

Figure 3.3: **Global & local glycan structural context in learned representations enable state-of-the-art prediction of glycan properties and lectin-glycan interactions (A)** GlyBERT had higher taxonomic classification accuracy than SweetTalk and SweetNet at each taxonomic rank. **(B & C)** GlyBERT demonstrated significantly improved predictive modeling of 20 lectins' glycan-binding preferences when compared to GlyNet and CCARL as seen in the parallel coordinates plot with median performances indicated by red bars (B) and in the bar plots with error bars indicating standard deviation across the 5-fold cross-validation (C). Significance was measured via a Friedman test followed by *post hoc* pairwise Wilcoxon signed-rank tests and Holm-Bonferroni correction (*: $p_{adj} < 0.05$, **: $p_{adj} < 0.01$, ***: $p_{adj} < 0.001$: ***).

subtree mining approach known as the Carbohydrate Classification Accounting for Restricted Linkages (CCARL) method [29]. All methods were trained and evaluated on the same set of 20 lectins with the same lectin-glycan interaction labels determined from microarray data from the Consortium for Functional Glycomics (CFG) [11] identically processed and divided into folds for training and cross-validation (CV) as provided by [29]. The glyBERT model trained above for glycan properties provided the basis for fine-tuning on the CFG microarray data, resulting in 20 separate models for the 20 lectins. Results for the other methods were previously reported [29, 22].

Performance was measured via area under the receiver operating characteristic curve (AUC), yielding a median AUC over the 20 lectins of 0.960 for glyBERT, demonstrating an increase in overall performance accuracy compared to GlyNet (0.946) and CCARL (0.896) with a statistically significant difference in ranked performance on matched lectins between glyBERT and the other two methods (Figure 3.3). The greatest increases in performance were seen for *Maackia amurensis* lectin II (MAL-II) and dendritic cell receptor DC-SIGN (Figure 3.3), indicating glycans on the CFG microarray likely manifested contextual dependencies. Indeed, DC-SIGN is known to have dual specificities for oligomannose glycans and blood group epitopes [58], with contextual oligomannose-preferences dependent on both the branch containing its primary motif and the mannose-content of surrounding branches [53]; furthermore, MAL-II has similar glycan preferences as MAL-I (Figure 3.1) but preferentially recognizes certain sialylated glycans when they are O-linked [55].

3.2.3. Deep generative exploration of the glycan structure-function space reveals contextual features discriminating immunogenic glycans

As illustrated in the preceding sections, machine learning models like glyBERT and others generalize structure-function relationships learned from known glycans in order to make predictions for new glycans. While to our knowledge no previous work has

gone further, such models also may also be used to *generate* new glycans according to unwritten specifications of desired structure-function relationships. This provides the opportunity to probe the vast diversity of glycan structures, aiding in hypothesis generation and informing directions for glycan-synthesis efforts. Additionally, it allows for *post hoc* analysis of captured patterns that could be distilled into more interpretable summarizations, e.g., by MotifFinder [81].

While full validation of computationally-generated glycans would require experimentation beyond the scope of the present study, we demonstrate here proof of concept by working with a subset of glycans that are structurally related but display different immunogenicity. We start with non-immunogenic glycans and apply a generative optimization approach, substituting monosaccharides at each step so as to push the glycans toward being immunogenic, based on probability gradients according to the glyBERT model. This enables an exploration of the glycan space near the starting non-immunogenic glycans, leading to some of those known be immunogenic along with novel glycans predicted to be immunogenic due to glyBERT's contextual assessment of their motifs. This process reached the known immunogenic counterparts within 2-3 steps (Figure 3.4) for four of the eight starting non-immunogenic glycans for which known immunogenic counterparts exist. For three of remaining four it arrived upon known, unlabeled glycans that exactly matched existing immunogenic glycans other than the anomeric configuration of the glycosidic bonds in the complete set of generative paths (Figure 3.6). In such a generative exploration, glycan structures that are generated more frequently (node size) might be more realistic (and here, more immunogenic) than others; notably the most frequently generated immunogenic structures in this case study were in fact known glycans. Generative steps that reverse direction to go against the probability gradient might indicate regions with few "nearby" realistic structures, as is likely the case for the non-realistic dead-end path

from NeuAc(α2-3)Gal(β1-3)GalNAc to GalNAc(α2-3)Gal(β1-3)GalNAc.

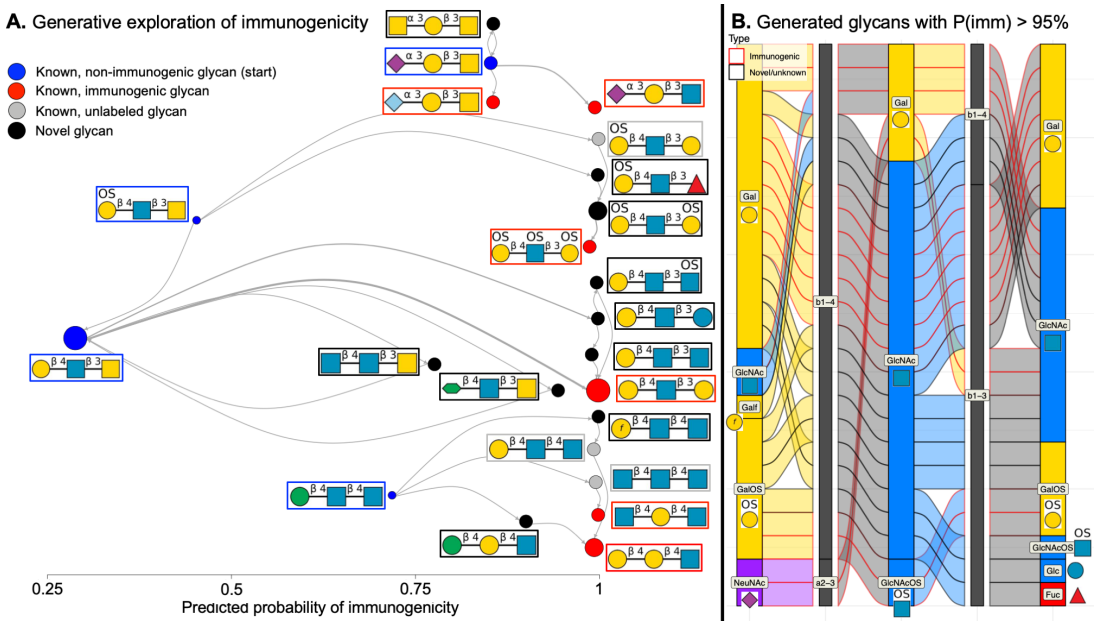

Figure 3.4: **Generative exploration of glycan structure-immunogenicity space (A)** Generative optimization starting from non-immunogenic glycans (blue) introduced monosaccharide substitutions to increase the probability of immunogenicity (x axis), allowing for the exploration of "nearby" known and unknown glycan structures (grey & black respectively), with some known glycans also labeled as immunogenic (red). Node sizes correspond to the number of times a glycan structure was reached along any of the generative paths, and edge widths correspond to the number of times a step between two glycan structures occurred across all the generative processes. Known immunogenic glycan structures were recovered for half of the starting, non-immunogenic glycans in only a few generative steps, demonstrating the process is able to generate realistic glycan structures and can optimize the generated structures for properties of interest. **(B)** This case study revealed features of glycans determined by glyBERT to confer immunogenicity in this structural context as shown in this alluvial plot [19] summarizing the structures of generated glycans with predicted probability exceeding 95%, including Gal and Gal derivatives in the non-reducing-end terminal position, GlcNAc in the middle position, and GlcNAc or Gal in the reducing-end terminal position.

Features in glycan structures deemed by glyBERT to be critical for immunogenicity are revealed by the generated glycans with the highest predicted probabilities of immunogenicity (Figure 3.4). While Gal and Gal derivatives at the non-reducing-end terminal position and GlcNAc monosaccharides at the middle position generally dom-

inate the generated glycans predicted to be immunogenic, the two non-immunogenic glycans with the lowest probability of immunogenicity share this monosaccharide composition at these positions, indicating the learned significance of GlcNAc and Gal monosaccharides but specifically not GalNAc at the reducing-end terminal position in this context. The substitution of NeuAc for the generally immunogenic NeuGc monosaccharide [179] imparted less gain in predicted immunogenicity than substituting the reducing-end terminal GalNAc for GlcNAc in the generative paths shown at the top of Figure 3.4. These trends were generally recapitulated in the complete set of generative paths (Figure 3.6).

Section 3.3

Discussion

As experimental techniques evolve to enable richer studies of complex glycan structures, analytical techniques to study structure-function relationships must also evolve to effectively capture complex and context-dependent associations between glycan properties and the global contexts in which functional motifs and epitopes are presented. To suitably learn and leverage context-dependent glycan structure-function relationships, we developed an attention-based model (glyBERT) using minimally processed glycan encodings. Building on the pioneering efforts of [14] and [20], we further demonstrated the value of deep learning in glycobiology, here by representing, leveraging, and revealing contextual glycan features. We showed that the learned glycan representations are sensitive to both local glycan features and global contexts in relation to a variety of glycan functions and labels, that they support state-of-the-art predictive performance of glycan properties from structures, and that they enable exploration of structure-function relationships while generating novel glycans.

One limitation of the glyBERT approach was the lack of positional information

for monosaccharide modifications in the training data from the SugarBase database; this limitation manifested for example in the relatively poor prediction performance for lectin ABA in Figure 3.3 & Table 3.1. In general, distinguishing monosaccharides and modifications would allow the model to more easily learn relationships between rare monosaccharides and rare modifications using independent occurrences instead of relying on co-occurrence. And while glyBERT learned implicit relationships between chemically similar monosaccharides (Figure 3.2), additional pre-training based on stereochemical structure and composition of monosaccharides and their modifications could further improve generalizability to more rare residues.

While the generative case study presented here is limited in scope due to our desire to recapitulate some known, closely-related immunogenic/non-immunogenic glycan pairs, the same framework could be applied to any property of interest and the gradient-based sampling method could be expanded to explore much more drastic changes to glycan structures. These expansions would truly enable deep glycan structure-function space exploration and contribute contextual-dependencies to glycoengineering efforts that might be applied to the design of competitive inhibitors or glycomemetics.

Section 3.4

Data and Code Availability

The glycan structures and associated labels used in this study were retrieved from SugarBase v2.0 (`https://webapps.wyss.harvard.edu/sugarbase/`) [13]. Lectin-glycan interaction data were retrieved from the supplemental information of [29]. All code and data used to perform this analysis, including training/testing splits and class labels, is available at `https://github.com/demattox/glyBERT`.

Section 3.5

Methods and Materials

A total of 16,048 unique glycan structures were encoded to an array-format compatible with glyBERT while maintaining all relevant chemical and branching information (Figure 3.5, subsection 3.5.3). These glycans were split into a training and test set following the same 80/20 split strategy as SweetNet [20] although the exact split used in that study was not available for direct comparison. Within the training set, glyBERT was pre-trained by predicting masked pieces of these structures as well as relevant labels (Figure 3.5, subsection 3.5.5). For each prediction task with reported performance, the pre-trained glyBERT model was fine-tuned by training specifically for each individual task, allowing the model to retain the most relevant information learned in the pre-training while focusing more specifically on each different task separately (Figure 3.5, subsection 3.5.6). Generative exploration was performed by using the immunogenicity prediction layer to calculate probability gradients and substitute individual monosaccharides that would lead to glycan structures with higher overall predicted probability of immunogenicity (Figure 3.5, subsection 3.5.7).

3.5.1. Data collection and cleaning – SugarBase glycans

From the 19,299 available glycan entries in the SugarBase v2 database [13] downloaded on November 11th 2020, 1,145 entries were eliminated because the glycan IUPAC was improperly or unexpectedly formatted. This included ambiguous or unmatched parentheses/brackets, incomplete orphaned glycosidic bonds, and branching brackets around the first monosaccharide or group of monosaccharides in the IUPAC, indicating that a portion of the glycan might be missing (e.g., SBID 712: **[Glc(a1-6)]4)**Glc(a1-4)Glc(a1-4) ... Glc(a1-4)Glc). An additional 1,319 entries were eliminated after discovering they were duplicates of glycans stored in other entries

in SugarBase with the order of the branches rearranged in their IUPAC representations (e.g., from SBID8088 & SBID13110: GalNAc(b1-4)[NeuNAc(a2-3)]Gal & NeuNAc(a2-3)[GalNAc(b1-4)]Gal). To facilitate systematic representation of branching, one glycan was omitted as an outlier with 25 branches since the next highest observed number of branches in a glycan was 11. Lastly, 786 glycans were omitted with any of the 531 very rare monosaccharides occurring in fewer than four glycans as there were not enough examples of these monosaccharides to allow for sufficient pre-training. Of the 19,299 original entries, we used the 16,048 unique glycans with sufficiently common monosaccharide vocabularies for the remainder of this work.

Taxonomy labels at varied taxonomic rank were recovered from the provided lists of species of origin for each SugarBase glycan entry by searching for the provided genus-level taxonomic ID (taxid) in the Python implementation of the NCBI's taxonomy tool Ete3 v3.1.2 [68] (local data written on December 29, 2020). Sugarbase provided 835 unique genera, and all but 15 were successfully matched to NCBI taxids. Of the 15 missing genera, 11 were viruses without scientific names; these were manually added to the "Virus" superkingdom/domain. The remaining 4 were from a typo or cases of deprecated or unofficial genus classifications and were manually corrected (*Columbia livia* → *Columba livia*, *Kinetoplastids* → *Kinetoplastida*, *Actinogyra* → *Umbilicaria*, *Arecastrum* → *Syagrus*). Some species/genera did not have defined taxids at certain taxonomic ranks. As taxonomic origin was not the primary goal of our study, this was manually addressed at the kingdom level only since NCBI taxonomy only categorizes eukaryota at the kingdom level and there were the most missing labels at this rank. To recover some labels for a significant portion of the glycans at this rank, "Bacteria", "Virus", & "Archaea" labels were carried over from the superkingdom/domain rank to the kingdom rank. Beyond the aforementioned adjustments, only the NCBI defined and provided taxids were used in all other cases

to ensure the quality and reproduciblity of labels used to train the model.

Labels were rectified over replicated glycans as follows. For linkage or immunogenicity labels from replicated glycans, the label was set to "None"/"Unknown" when directly conflicting labels were present, or the definitive label was used if other replicates for a glycan were unlabelled. For species of origin labels from replicated glycans, the set of all observed species of origin from all replicates of the glycan was used.

3.5.2. Data collection and cleaning – CFG glycans

Unique glycans from the Consortium for Functional Glycomics (CFG) microarray v5.0 were processed to match the specific IUPAC formatting utilized by SugarBase. To ensure patterns and relationships learned from training on SugarBase glycans could be leveraged, monosaccharide names from the CFG glycans were manually adjusted to match the names used in SugarBase glycans, including the reordering of modifications (GlcNA $\rightarrow$ GlcAN) and the removal of position specific information from sulfation modifications ((3S)Gal / (6S)Gal / (3S)(6S)Gal $\rightarrow$ GalOS).

Lectin binding to CFG microarray glycans was determined from data provided in supplemental file 6 of [29], splitting binding interactions into the bound class or the unbound class based on the Median Absolute Deviation (MAD) technique, while discarding intermediate interactions to maximize type I and type II error control.

3.5.3. Glycan encoding

Glycan IUPAC strings were processed into custom tree objects built from anytree v2.8.0 [5], placing each monosaccharide into a node while recording glycosidic linkage information and anomeric conformation (Figure 3.5). A "START" token was placed before the reducing-end terminal monosaccharide (root node) and "END" tokens were added after each non-reducing-end terminal monosaccharide. The "START" token allowed for embedded representations of entire whole glycans and the "END" token

was designed to allow for more dramatic changes to the glycan structure in future generative processes. These tree structures were then encoded into seven-column arrays with each node placed into its own row, recording the following information for each saccharide in its columns: (1) identity, represented by a single number corresponding to the rank order of the unique monosaccharides by frequency; (2) anomeric conformation, α, β, or Unknown; (3) indices of carbons involved in glycosidic bonds to other monosaccharides; (4) position within the branched structure glycan in a "subway line" approach described in the following paragraph; (5) index of the carbon linking to the parent monosaccharide; (6) depth, or distance from the root node; (7) index in a list of monosaccharides at the same depth on different branches, ordered by following the lowest carbon number at each branch point in a depth-first tree traversal. The carbons involved in glycosidic bonds to other monosaccharides were stored as 9-bit binary numbers, with each position representing a carbon number for a given monosaccharide participating in a glycosidic bond set to 1 and the remaining positions set to 0.

Glycan branching was explicitly accounted for in this encoding with the "subway line" approach illustrated in Figure 3.5, where lines were drawn along monosaccharide nodes from each non-reducing-end terminal monosaccharide to the reducing-end terminal monosaccharide (root node) resulting in what appears like a map of subway lines that account for each branching event in a glycan. For each monosaccharide, the branches (or subway lines) that pass through that node (or subway station) were recorded in an 11-bit binary number, with the bits corresponding to branches passing through the node set to 1. An 11-bit binary number was used to have a consistent branching representation for all glycans with up to 11 branches. The 9-bit glycosidic bond binaries and 11-bit branch position binary numbers were converted into decimal numbers in the final encoding to simplify representation and reduce dimensionality.

Depicted glycan representations follow the SNFG system [160] and were rendered using DrawGlycan-SNFG [26].

3.5.4. Architecture and Implementation

GlyBERT was implemented in Pytorch [123] & Fairseq [120] following the architecture diagrammed in Figure 3.5. For embedding layers, the original BERT architecture embedded all different input feature encodings to the same dimension and combined them, adding positional (index of word) and segment (context information) embeddings to token embeddings. To adapt this approach to glycans, the monosaccharide encodings described above (anomeric state, carbon linkage, branch, parent, and depth) were embedded and added to the monosaccharide embedding. Learnable embedding layers were utilized for these different types of encoding features. For the transformer architecture, we used the same 12 multi-head self-attention layers (each with 12 heads) as the $\text{BERT}_{\text{base}}$ model. The classifier layers were composed of 6 Multilayer Perceptrons (MLPs) with one for each of our customized pre-training tasks. GlyBERT was optimized with Adam [79].

3.5.5. Pre-training

In order to begin building representations of the glycans that capture the structural diversity of all available glycans from SugarBase with contextual relevance to biologically relevant properties, we employed a pre-training process inspired by the original BERT work [39]. We used six pre-training tasks, some semi-supervised and some supervised:

Semi-supervised To build an implicit understanding of the different monosaccharides and their glycosidic bonds and their use in glycans in different contexts, pieces of each glycan structure (monosaccharide identity, anomeric conformation, and gly-

cosidic bond information) from the the training set were masked, and the model was trained to predict the masked information based on its surrounding local and global context in the glycan structure. This approach treats monosaccharide types as words connected in branched sentences and follows pre-training strategies of conventional language models. All three masked prediction tasks followed the "mask LM" (MLM) procedure [155], here masking 25% of each glycan input (before padding) independently by a "MASK" tag. Only the cross entropy loss of masked tokens was computed, following [39].

Supervised To incorporate biologically relevant information into the glycan representations, supervised tasks included predicting N/O glycoprotein linkage, immunogenicity, and taxonomic origin. MLPs for these tasks took the embeddings of the "START" tokens as representative inputs for the whole glycans. Cross entropy loss was used for N/O linkage and immunogenicity prediction. As glycans can be found in multiple organisms, taxonomic origin prediction was treated as a multi-class, multi-label prediction task and binary cross-entropy was employed.

3.5.6. Fine-tuning

GlyBERT was fine-tuned separately to study each specific glycan property for which predictive performance was reported (Figure 3.5). In this step the other MLPs' prediction heads were simply disabled.

For the downstream fine-tuning task of predicting binding information on CFG data, we used the pre-trained model without its classification layers. A 3-layer MLP, taking the "START" token embedding as input, was then appended to the pre-trained model. Since one model was trained for each lectin's binding activity, we used binary cross entropy loss. Models were trained using the same stratified 5-fold cross-validation as GlyNet and CCARL and optimized with Adam [79].

3.5.7. Generative optimization

To optimize glycans for a given property, two functions are needed: a differentiable function that is able to evaluate how well the current glycan meets the desired property and a function to map glycan representations in the latent space back to the encoding space [56]. In the glyBERT architecture, the prediction layer serves as the property evaluation function; since it is a MLP that combines a sequence of linear operations and differentiable non-linear activation functions, it is able to predict the property of interest and is differentiable. The MLM layer serves as the function to map a glycan's latent space representation back to a probability distribution belonging to the encoding space.

With these two functions in hand, the pipeline for generating glycans follows naturally (Figure 3.5), based on the embedding, attention, MLM, and prediction layers from the trained glyBERT model. For a given glycan structure used as the starting point, its encoding was input into the embedding layer, returning an initial embedding. The multi-head attention layer provided the glycan representation and monosaccharide token embeddings in the latent space. Monosaccharide embeddings were used to compute the original monosaccharide type probability distribution, M_0, by the MLM layer. The glycan representation was then passed through the immunogenicity label prediction layer (serving as the property evaluation function) and a cross entropy loss against the positive (immunogenic) label was computed.

To update the embedding in the latent space, the network parameters were frozen and the gradient of loss with respect to the embedding was calculated. The embedding was then updated by the gradient with a learning rate of 0.1, thereby computing the optimized monosaccharide type distribution, M_1, for each token.

The original and optimized monosaccharide type probability distributions, M_0 and M_1, were used to pick which token to change and which monosaccharide to

change it to. Two vectors, P_0 and P_1, were generated from M_0 and M_1, giving the probabilities of each token being the current monosaccharide type. These vectors were then used to calculate a discrete probability distribution M_s from which to select the monosaccharide to change, where:

$$M_s = \frac{abs(P_0 - P_1)}{||P_0 - P_1||_1} \tag{3.1}$$

For the i^{th} token sampled from M_s, P_s is the probability distribution of the i^{th} token's monosaccharide type. To ensure a different glycan was returned after each round of generation, the probability of being the original monosaccharide was changed to 0 in P_s, with the altered P_s referred to as P'_s. Finally, the new monosaccharide was sampled from $\frac{P'_s}{||P'_s||_1}$ and placed into the glycan structure. To avoid using potentially unrealistic glycosidic bonds while calculating the gradient, once the new monosaccharide was selected, the anomeric conformations and carbon linkages encoding of substituted monosaccharides were left unspecified (using the same "mask" token to mark them as such). Since any context (including glycosidic bond configuration) can impact glycan functionality, it was preferred to calculate the gradient for the next substitution using only verified features of the glycan structures. In future implementations, this might be addressed instead by expanded generative processes to sample updated glycosidic bonds simultaneously, or potentially filtering out glycans unable to be synthesized in an organism of interest [110] depending on the intended use case.

This process was iterated for each starting glycan, providing a series of glycans that generally had increasing probabilities of being immunogenic until a known immunogenic glycan was found or a fixed number of structures (10 in the presented results) were generated (Figure 3.5).

The validation set of glycans for the generative process were taken from sets of matching glycan structures (ignoring monosaccharide identities) where each set

contained a sufficient number of immunogenic and non-immunogenic pairs. Of the 20 sets of matched glycans structures with at least one immunogenic and one non-immunogenic glycan, 4 were selected for which there were sufficient positive and negative examples (more than 5) and immunogenicity classification was strong ($> 80\%$ accuracy), thus providing higher confidence in the interpretation of results. From these 4 sets, there were 14 immunogenic and 8 non-immunogenic glycans in total.

Section 3.6

Supplemental

CFG v5.0 Probe Num.	Probe	RFU (100 μg ABA)	MAD binding
25	(3S)Galb1-4Glcb-Sp8	1066	Bound
42	(6S)Galb1-4Glcb-Sp0	34	Not bound
43	(6S)Galb1-4Glcb-Sp8	37	Not bound

Table 3.1: ABA preferentially binds a glycan with an OS modification on the terminal galactose but does not bind the same glycan when the OS modification is present on C-6. All three glycans are represented in our encoding as GalOS(b1-4)Glcb based on the modification-position agnostic monosaccharide vocabulary found in SugarBase. RFU values are reported from the CFG mammalian microarray v5.0 data used by [29] with ABA at 100 μg available from `http://www.functionalglycomics.org:80/glycomics/HFileServlet?operation=downloadRawFile&fileType=DAT&sideMenu=no&objId=1004226`. The binding label determined by [29] via the MAD technique is reported in the 4^{th} column.

Section 3.7

Funding

This work was supported in part by the Burroughs Wellcome Fund Big Data in the Life Sciences training grant awarded to DEM, and NIH R01 2R01GM098977 to CBK.

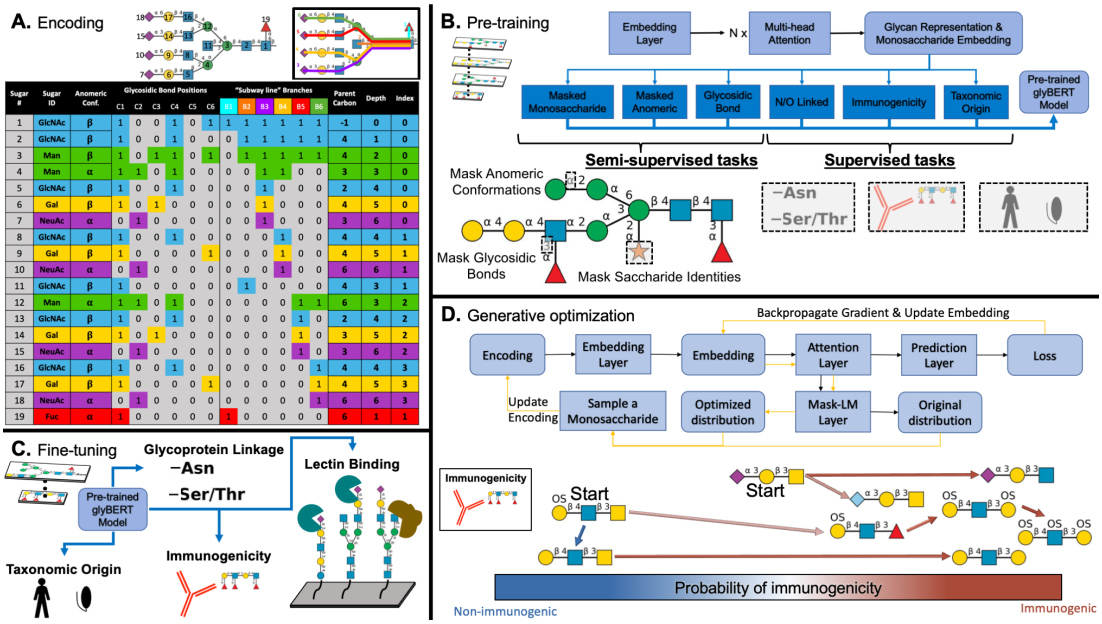

Figure 3.5: **Overview of glycan encoding and glyBERT pre-training, fine-tuning, & generative exploration.** **(A)** A simplified example of the glycan encoding used here with fewer bits and without the "START" and "STOP" tokens. Each monosaccharide in a glycan is fully represented in its own row, with an explicit representation of branching using the "subway line" approach demonstrated in the inset in the top right of the panel. The indicator variable columns used to indicate the involvement of carbons in glycosidic bonds or the "subway lines" that pass through each monosaccharide are treated as binary numbers and collapsed into decimal numbers to simplify encoding representations for the model. **(B)** The model architecture and prediction tasks used in pre-training glyBERT. **(C)** Fine-tuning of glyBERT to build specialized glycan representations specific for each prediction task. **(D)** The model architecture used in the generative process. For each generative iteration, there are two components represented by black and yellow arrows. Black arrows indicate the first component, with the goal of getting the status of the current glycan structure before making any changes or updates. From the original encoding, simply passing through the embedding, attention, and prediction layers yield the current monosaccharide embeddings and the current loss towards the optimization target. Yellow arrows indicate the second component, optimization and updating. Embeddings are updated by utilizing the calculated gradient to get new probability distributions for each potential monosaccharide substituent. The final discrete update in the encoding is then generated by random sampling from these probability distributions.

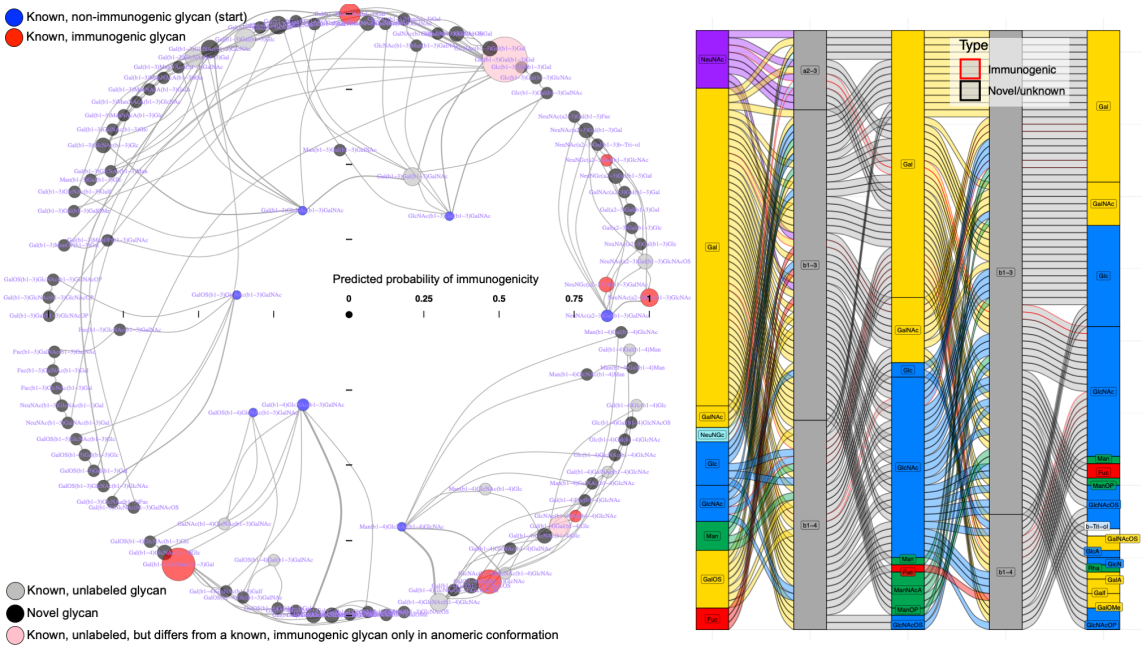

Figure 3.6: **Complete generative exploration of glycans' structural relationship to immunogenicity.** Generative paths from seven non-immunogenic glycans, including the four in Figure 3.4 along with three additional glycans from which known glycans were discovered which were unlabeled but shared the same carbon attachments and monosaccharide identities as known immunogenic glycans with differing anomeric conformations. In general, the glycans generated here showed similar insights as Figure 3.4 into the features driving glyBERT's prediction glycan immunogenicity in this structural context: Gal & Gal derivatives were most common at the non-reducing-end terminal position with GlcNAc or Gal monosaccharides appearing frequently in the middle position, although Gal & GalNAc are better represented here compared to Figure 3.4. The largest apparent difference from this generative exploration compared to Figure 3.4 is the presence of glycans with reducing-end terminal GalNAc with high predicted probability of immunogenicity, highlighting other drivers of immunogenicity learned by glyBERT relevant in this structural context.

Section 3.8

Acknowledgments

We would like to thank Zach Klamer for discussion while the manuscript was being prepared, as well as Professors Margaret Ackerman, Karl Griswold, and Jiwon Lee of the Dartmouth Thayer School of Engineering, and their lab members, for their continued feedback on this project and many others at all stages of development.

Section 3.9

Author contributions

BD Conceptualization (Equal), Writing – Original Draft Preparation (Equal), Writing – Review & Editing (Equal), Methodology (Lead), Software (Equal), Investigation (Equal), Formal Analysis (Equal), Visualization (Equal)

DEM Conceptualization (Equal), Writing – Original Draft Preparation (Equal), Writing – Review & Editing (Equal), Methodology (Equal), Software (Equal), Investigation (Equal), Formal Analysis (Equal), Visualization (Lead), Data Curation (Lead), Funding Acquisition (Support)

CBK Conceptualization (Equal), Writing – Original Draft Preparation (Supporting), Writing – Review & Editing (Equal), Methodology (Supporting), Project administration (Lead), Visualization (Equal), Funding Acquisition (Lead)

Chapter 4

Protein Design Algorithms for Evasion and Mapping of Antigen Recognition by Uncharacterized Antibodies

The ability of antibodies (Abs) to sensitively and specifically recognize particular regions, or epitopes, on their antigens (Ags) is fundamental to the immune response against pathogens and is also key to the development of wide-ranging biotherapeutics and vaccines. In some therapeutic applications, it is desirable to modify a target protein to evade particular Abs, e.g., to enable a biotherapeutic to evade preexisting anti-drug antibodies or to focus a vaccine response on a specific neutralizing epitope. Ag variants that evade specific Ab recognition can also serve as reagents to sort Abs by their epitopes and even localize those epitopes on the Ag, according to which Ag variants disrupt which Ab binding. In redesigning an Ag to evade a specified Ab interaction, mutations must be chosen to balance the effects of candidate mutations on Ab recognition and Ag stability and function. Furthermore, in many applications,

there are multiple Abs to evade, some or all of which may not have been characterized. We here develop geometry-based algorithms to design Ags to evade preexisting Abs (known or unknown), aka "B cell epitope deletion" and to design panels of Ags to enable simultaneous separation and epitope localization, which we term "epitope map deconvolution", of multiple (known or unknown) Abs in a polyclonal mixture. We first validated our hypothesis that a representative set of Abs, presumably not directly related to the Abs being characterized, provides a sufficient basis by which to predict the actual epitopes of the actual Abs. Then in retrospective case studies, we demonstrated that the algorithms manifests promising recall and robustness of designing suitable Ab-evading Ag variants. For B cell epitope deletion, our algorithm identified experimentally validated epitope-deleting variants as well as additional promising mutation options. In epitope map deconvolution, we showed that our designed panels can be expected to distinguish different Abs according to their different epitope specificities better than would be possible with more traditional mutational approaches. Our work represents the first methods to systematically design Ag variants for characterization and evasion of polyclonal Ab responses, even in the absence of any prior characterization of the Abs.

Section 4.1

Introduction

The adaptive immune response against a pathogen often culminates with the development of B cells selected and matured for potent recognition of pathogen Ags [143]. The Abs produced by these B cells can in some cases directly neutralize the target pathogen by binding and blocking an important functional site on a key Ag (e.g., mediating cell receptor recognition [71, 47]), and in other cases their overall binding and opsonization of an Ag can guide immune-mediated clearance (e.g., non-neutralizing

antibodies contribute to HIV's immune escape [87]). High-affinity and -specificity Ab-Ag binding is mediated by recognition between particular interfaces on their surfaces, namely the epitope of the Ag and the paratope of the Ab. B cells share a common set of germline genes from which Abs are developed [4], and Ag-binding fragments (Fabs) share a common overall structure, comprised of relatively unchanged framework regions providing a scaffold and relatively variable complementarity determining regions (CDRs) defining the paratope and more directly interacting with the Ag. Diversity in paratope residues is the main source of differentiation in Ab targeting of different Ags, though in some cases particular framework mutations have also been shown to have significant impact [143]. The processes of VD(J) recombination and somatic hypermutation introduce extensive diversity to the paratopes (and to a lesser extent in the frameworks), enabling the maturation process to generate a potent Ab from a relatively weak-binding, non-specific progenitor [115].

When the Ag being recognized by the immune system is actually a therapeutic, it can be beneficial to modify it so as to diminish or even entirely escape Ab recognition. For example, if a biotherapeutic is derived from a non-self source that a patient may have naturally experienced, then that patient's immune systems may have already developed Abs against it, i.e., anti-drug antibodies (ADAs), which may reduce biotherapeutic efficacy or cause even worse reactions [130]. A prominent example is the 38-kDa fragment of *Pseudomonas* exotoxin A (PE38), which is part of a recombinant immunotoxin designed to treat cancer, but elicits a potent ADA response due to its origins in a common pathogen [122]. Engineering a biotherapeutic to avoid preexisting Abs is often called "B cell epitope deletion" as it requires modifying the protein so that it no longer presents the epitopes against which existing Abs were matured [57, 109, 138]. We note that avoiding the initial formation of an ADA response in a naive patient, or a new ADA response against a B cell epitope-deleted

variant, requires deletion of T cell epitopes that are essential for T cell help to drive the development of potent class-switched Abs; that is a separate problem [57]. To address the preexisting ADA problem with PE38, [94] employed scanning mutagenesis, mutating known epitope residues to alanine or glycine and evaluating effects on Ab binding, ultimately deriving an engineered variant that showed low reactivity with Ab serum but maintained high antitumor activity. Similarly, [139] deimmunized truncated diphtheria-toxin (DT) by mutating hydrophilic surface residues to amino acids randomly chosen from glycine, alanine, and serine. [59] used random mutagenesis to reduce the interaction of antibodies against Chemotaxis inhibitory protein of *Staphylococcus aureus*.

Modifying an Ag to escape recognition of some Abs has also been used to design vaccine Ags that yield a more focused immune response against an epitope of interest, preserving that epitope while modifying other parts of the Ag to reduce recognition of undesired epitopes [182]. The enormous diversity of HIV-1 variants, due to its rapid and extensive evolution, makes vaccine development (as well as of course natural immunity) extremely difficult, since Abs elicited against one variant might not see their desired epitope on another variant, and only a small fraction of binding Abs can broadly neutralize the virus [121, 104]. To discover broadly neutralizing Abs (bnAbs) that bind at an important conserved and neutralizing region, [173] resurfaced HIV-1 glycoprotein 120 (gp120) to obtain a molecule that preserved the CD4-binding site (CD4bs) antigenic area and eliminated all other antigenic areas, thereby providing a reagent with which to isolate CD4bs bnAbs. Dengue also presents a particularly important target for epitope focusing, as non-neutralizing, cross-reactive Abs against undesired epitopes can actually lead to severe disease (Ab-directed enhancement). Consequently, [133] engineered epitope-focused variants of Dengue virus envelope protein domain III (DENV DIII) using a combinatorial mutagenesis approach, obtaining

DENV variants with unwanted epitopes ablated.

Ab escape from Ag variants can also be directly used to characterize unknown epitopes. In fact, alanine scanning mutagenesis was developed for the purpose of determining binding sites [32, 112, 105]: one-by-one, mutate each residue to Ala, evaluate binding, and see which mutations ablate binding. Other approaches [66, 142] use computational methods to focus experimental efforts on relatively fewer but potentially more informative mutations. While these approaches to "mapping" epitopes has had numerous successful applications for monoclonal Abs [175, 96, 129], in some cases there is a set, or even a polyclonal mixture, of different Abs, which may recognize different epitopes. In a prominent example of identifying epitope residues for multiple Abs simultaneously (though not deconvolving which Abs recognize which residues), [118] evaluated effects on binding of alanine and glycine mutations at 41 highly exposed surface residues, evenly spread out over PE38, and thereby determined which positions were in the binding sites of a set of mice Abs. Experimental competition, or epitope binning, assays [18, 149, 74, 17] can help distinguish which Abs recognize overlapping epitopes, but don't directly localize those epitopes on the Ag. Recent computational-experimental methods [17, 98] simultaneously cluster and localize epitopes using a combination of computational docking and experimental competition assays, across different Ag variants. When confronted with a polyclonal mixture, Ag variants can be used to sort the Abs based on their epitopes, and potentially even help localize those epitopes. For example, [25] designed a panel of gp120 single mutation variants to map and sort CD4bs antibodies from polyclonal sera. We generalize these various related epitope characterization tasks into a problem we term "epitope map deconvolution": for a set of uncharacterized Abs, the goal is to design a panel of Ag variants that can enable separation of the Abs by their specificity and localization of their associated epitopes on the Ag.

We here develop algorithms to design Ag variants to escape recognition from a set of potentially uncharacterized Abs. Our B cell epitope deletion design algorithm optimizes individual variants broadly resurfaced at sites considered likely to be epitopes for preexisting Abs, seeking to disrupt their recognition while maintaining Ag stability. Our epitope map deconvolution design algorithm optimizes panels of variants resurfaced in different potential epitope patches so as to enable separating and localizing Abs and their epitopes. Through retrospective studies we demonstrate that in both cases, even when lacking any characterization of the associated Abs, our computational design methods are able to leverage general information regarding Ab-Ag recognition encoded in computational models in order to design effective escape variants.

Section 4.2

Results

Fig. 4.1 shows a schematic overview of the design algorithms. Since the true Abs may be unknown, we base the methods on a set of representative Abs, and demonstrate in 4.2.1 that these suffice to predict the actual epitopes of the actual Abs. For B cell epitope deletion, the sampling algorithm seeks to spread the targeted positions over the surface, selecting mutations for an Ag variant so as to disrupt potential Ab binding to important sites (and covering nearby sites) while not sacrificing Ag stability according to a predictor. We demonstrate in 4.2.2 that the designs include experimentally validated deimmunizing mutations, along with other plausible alternatives. For epitope map deconvolution, the design focuses on patches of nearby epitope residues, such that mutating them is likely to disrupt binding to a particular epitope and thereby enable identification and localization of Abs against it. The panel is designed by sampling patches across the surface (one per variant) and sampling binding-disruptive by

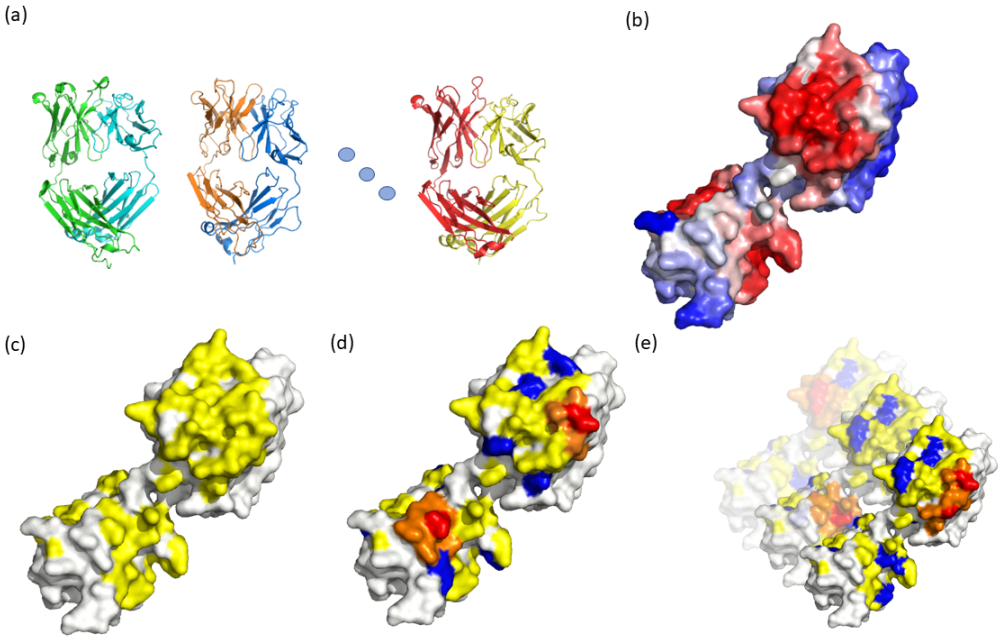

Figure 4.1: Key steps in the design of individual Ag variants for B cell epitopes deletion and panels of Ag variants for epitope map deconvolution. **(a)** Assuming that the true Abs are uncharacterized, a non-redundant set of representative Abs is used as "proxies", under the hypotheses that their paratopes will be sufficiently similar for the purposes of predicting the most likely epitopes on the Ag. **(b)** Epitopes are predicted for the representative Abs, and the union (max probability over all Abs) is taken, here with red representing high probability of being an epitope and blue representing low probability. **(c)** To focus the design problem, the residues to target (yellow) are selected by thresholding the predicted epitope probability (here at 0.5), identifying those residues where disruptive mutations are most likely to matter. **(d)** B cell epitope deleted variants are designed by geometric sampling of the targeted residues. Each design includes a set of mutations that are predicted to be disruptive to Ab binding but not to Ag stability, according to predictive models. Positions are chosen to cover a large fraction of the other targeted residues, in that the selected mutations are also likely to disrupt binding at these nearby positions. The example shows some chosen positions in red, with the nearby epitopes they cover in orange; for clarity the covered epitopes aren't shown for the remainder of the chosen positions (blue). **(e)** An epitope map deconvolution panel is designed by geometric sampling of the targeted patches following by residues within the sampled patches. Each design includes a panel of variants that will disrupt binding at different locations. Mutations (red) of the same variant are chosen to cover residues within the same patch (orange).

stability-preserving mutations within the patches. We demonstrate in 4.2.3 that the designs are able to distinguish different Abs targeted on the same Ag at different

epitopes.

4.2.1. Representative Antibodies

In order to computationally design Ag variants to evade recognition by a set of Abs (as opposed to, e.g., entirely resurfacing the Ag, or experimentally generating and testing variants), it is necessary to predict which epitope residues are being recognized by the Abs. One approach would be to perform Ab-independent prediction, essentially identifying solvent-exposed "knobby" residues that are generally recognizable by Abs. However, when the Ab is known, Ab-specific epitope prediction achieves state-of-the-art prediction performance [33], since the paratope provides valuable information about what parts of the Ag actually accommodate binding by that particular Ab. We hypothesize that a set of already-characterized Abs can sufficiently represent the paratope diversity manifested by the set of true Abs being targeted, so that the benefits of Ab-specific prediction can be leveraged to obtain higher predictive performance than Ab-independent prediction. We thus selected a set of 41 representative Abs clustered from SabDAb [42] to 60% sequence identity, and used the state-of-the-art epitope predictor PInet [33] to make Ab-specific predictions, combining predictions across the representatives.

We tested this "representative Ab hypothesis" by comparing a predictor based on PInet and representative Abs against two state-of-the-art Ab-independent epitope predictors Discotope 2.0 [85] and SEPPA 3.0 [186]. As a benchmark we used the same non-redundant set of 41 Ab-Ag pairs from SAbDab, for each Ag omitting the partner Ab from the set of representative Abs used in our method. We evaluate both recall (the ratio between correctly predicted epitope residues and all epitope residues) and precision (the ratio between correctly predicted epitope residues and all residues predicted to be epitopes). For design of Ab-evading variants, it is most important to identify most of the true epitope residues, even at the expense of including some

non-epitope residues; i.e., recall is paramount.

Table 4.1 summarizes the prediction results. Discotope and SEPPA are able to achieve relatively higher precision, but prediction using PInet and representative Abs results in substantially higher recall, our goal. We also explored the impact of the size of the representative Abs set. As shown in Fig. 4.2, the precision and recall of epitope prediction suffer if the set of representative Abs is too small, so that there is insufficient paratope diversity in the set to represent the true paratope. However, even with a relatively modest number of 10 representative Abs, the performance is close to that achieved with the full set of 41 or even the true Ab, though with more poorly-predicted outliers. It remains to be seen how these results change with further improvement of epitope predictors, for which more refined differences in paratopes might make more difference in predicted epitopes.

Table 4.1: Ab-specific epitope prediction using PInet with a set of representative Abs achieves much higher recall than Ab-independent prediction with Discotope or SEPPA, at some loss of precision. PInet epitopes were predicted according to a 0.5 threshold while Discotope and SEPPA do not use thresholds.

Test	Model	Recall	Precision
	Discotope 2.0	0.29	**0.30**
	SEPPA 3.0	0.67	0.27
	PInet	**0.85**	0.20

4.2.2. B Cell Epitope Deletion

We used two different retrospective case studies, along with a larger-scale benchmark derived from SabDab, to evaluate our algorithm for designing B cell epitope deleted variants of a given therapeutic protein. While, as would be expected given the size of the available sequence space, many of the exact variants designed by our approach were not tested in the previous studies, the overlap and similarities support our confidence in the quality of the designs. These studies, along with the benchmark, further provide the opportunity to assess the robustness of our approach and trends

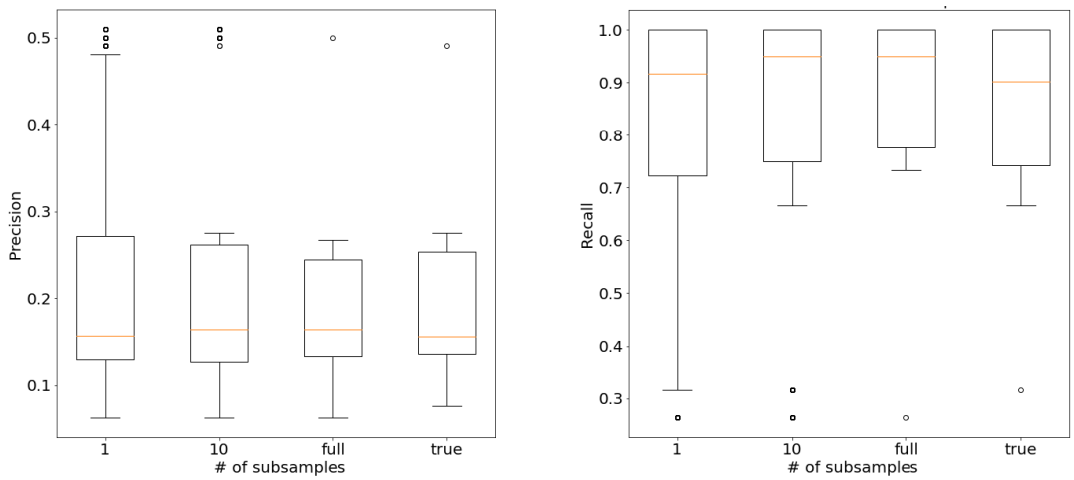

Figure 4.2: Epitope prediction performance (left: precision; right: recall) using representative Abs approaches that using the true Ab with the full set of 41 representative Abs or even a reduced set of 10, but falls off with only 1. Boxplots are taken over different random subsets of the full set, at the specified subset size. Though using one random Ab can still sometimes yield good predictions, performance is more stable with more Abs.

over mutational loads and other parameters.

EGFP. Our first case study involves deletion of a nanobody epitope from enhanced green fluorescent protein (EGFP), a target of previous epitope deletion efforts using the DisruPPI method [27], in which designs were based on computational optimization of trade-offs between predicted effects of mutations on EGFP stability and disruptiveness to the interaction according the complex structure (PDB ID 3OGO, chains B and G).

We first used our approach based on the same knowledge as DisruPPI, namely focusing design on the known nanobody epitope from the crystal structure, rather than relying on PInet and representative Abs. Table 4.2 lists the top 10 single mutation designs from our algorithm, which includes the two designs (N164K and R162E) previously experimentally demonstrated to be stable and disruptive. Additional variants also showed similar or better levels of epitope coverage (i.e., more of the other epitope

residues are nearby and thus Ab interactions involving them are more likely to also be perturbed) and predicted stability (i.e., the mutations are not likely to destabilize the Ag's structure), and thus would be candidates worth considering.

Table 4.2: Given the known epitope, our design algorithm identified two validated single-point mutations (red) while also providing additional options predicted to be similarly beneficial in terms of the fraction of the known epitope residues covered (i.e., sufficiently near the selected mutation sites) and the predicted stability of EGFP (as assessed by a protein language model).

Rank	Mutation	Coverage	Stability
1	N164K	0.938	0.938
2	E166K	0.807	0.882
3	D167G	0.807	0.838
4	A200V	0.576	0.782
5	S141E	0.576	0.617
6	Y137F	0.315	0.666
7	R162E	0.315	0.666
8	F217T	0.576	0.382
9	S169G	0.807	0.111
10	N138Q	0.807	0.085

We now turn to the case without the complex structure, applying the algorithm based on PInet predictions for the representative Abs. We designed variants with different mutational loads and different epitope probability thresholds for which residues to target (from the default 0.5 down to 0.0, i.e., all surface residues). We evaluated the coverage of epitope residues by the top design, or covered in total by the top several designs. Depending on the protein target and the experimental setup, the results from the top several designs may or may not provide information useful for comparison and integration. For example, in some cases, there may be sufficiently detectable disruption from each of several designs to follow up with new variants combining their mutations. In any case, the "union of top designs" coverage metric captures the relative diversity of the designs, since highly overlapping designs will provide the same coverage as just one of them. As the baseline, we generated random

designs of the same mutational loads being used in the designs, thereby assessing to what extent design improves over random expectation.

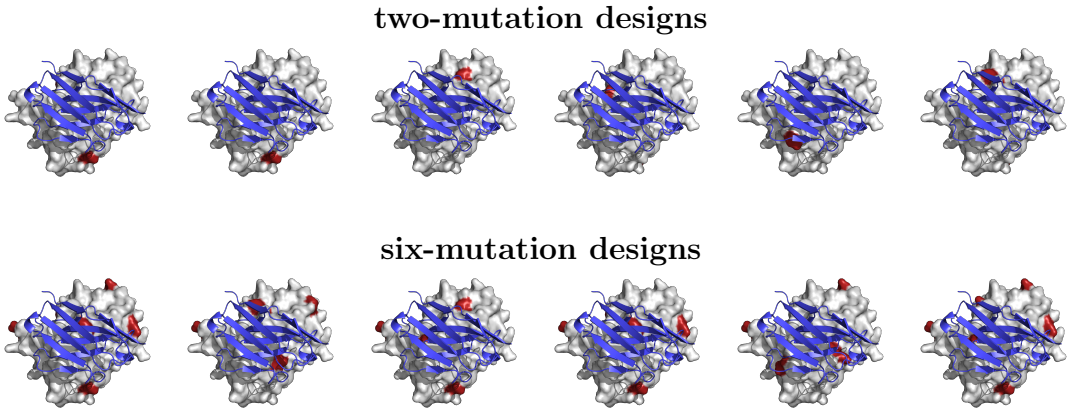

Figure 4.3: Top six 2-mutation (top row) and top six 6-mutation (bottom row) epitope-deleting EGFP designs (red: mutated sites; blue cartoon: nanobody, for illustration but not used in design).

As shown in Fig. 4.3(top), for each of the top six 2-mutation designs, there is only one mutation visible in the epitope region, yielding a low percent coverage (10%). This is because PInet predicted epitopes (Suppl. Fig. 4.10) on both sides of the protein and our geometric sampling algorithm to select positions to mutate maximizes distance between them. If we take the union of the residues comprising these designs, the total coverage is much higher (50%), as the designs are hitting diverse residues. When the mutational load is increased to six (Fig. 4.3(bottom)), the total coverage for the top design dramatically increases to 40%, and since the mutations are quite diverse, the union coverage for the top six designs reaches 90%.

Fig. 4.4 elaborates and quantifies these trends over additional mutational loads and different numbers of top designs. As would be expected, more mutations or more designs leads to better coverage, though even at moderate levels of mutations or designs, the coverage is already quite good; e.g., the single top 5-mutation design achieves 20% coverage, while the 6 top 2-mutation designs achieves in the range 40-90% coverage. At all but the lowest settings, the coverage is substantially higher

than would be expected at random, as illustrated by the boxplots. The figure also illustrates the effect of the epitope probability threshold, from the default 0.5 at the right of each line down to 0 at the left. While the trends of the threshold vary at different mutational loads and numbers of top designs, it seems that a threshold of 0.3 or 0.4 is generally best, presumably accepting more false positives in order to obtain more true positives.

While our basic approach is to take the top ranked designs, an alternative is to subselect from the designs a "diversity-filtered" set of variants, such that each introduces at least one mutation not seen in the previous ones. This naturally leads to better coverage, as illustrated on the structure in Suppl. Fig. 4.11 and in trends in Suppl. Fig. 4.12. These diversity-filtered designs substantially improve coverage at low mutational loads; e.g., the coverage from the union of the top six diversity-filtered 2-mutation designs is 80%, 30% higher than simply picking the top six 2-mutation designs.

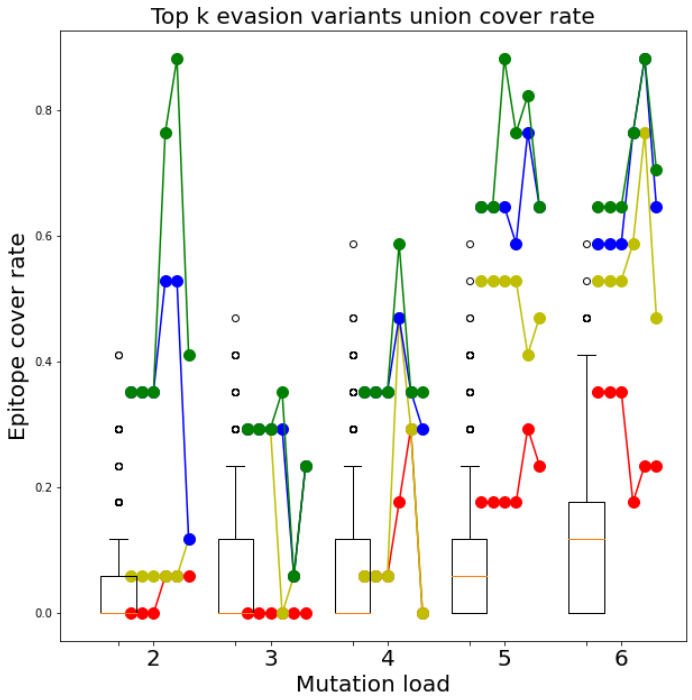

Figure 4.4: Coverage of known EGFP epitope residues by designs (lines) and random sets of mutations (boxplots) at different mutation loads (groups on the x axis). Designs are chosen at epitope prediction probability thresholds (points on the lines from 0.0 to 0.5) and the coverage over the top several designs (red: 1, yellow: 3, blue: 6, green: 9) is used to characterize design diversity.

PE38. We next applied the algorithm to PE38, which, as discussed in the introduction, is part of an immunotoxin that has been shown to have multiple B-cell epitopes and has been subjected to previous B cell epitope deletion efforts [118]. In contrast to the single nanobody against EGFP, for PE38 we need to disrupt binding sites from multiple (unknown) preexisting Abs, and thus employed PInet with representative Abs. We designed variants with mutational loads ranging from 2 to 10, stepping by 2.

Fig. 4.5 shows epitope coverage by the top three 6-mutation designs. Notably, most of the mutations that aren't at a known epitope position actually cover one (i.e.,

are within 12Å of it), and in fact for more than 60% of the epitope residues, there is at least one mutation from the top three designs covering that site. It is hard to make a direct comparison of our designed mutations to those previously reported [94], which were derived by targeted alanine scanning and tests of more than 50 variants, with 14 of those variant mutations affecting binding. Among these 14 mutations, two of the same positions (E431 and R490) were included in our top design, another (at K590) was in the top three, and six more were in lower rank designs.

As with EGFP, we quantified the performance of the top designs, over mutational loads and epitope thresholds, and compared against random designs. In addition to taking the union of the top designs, we also evaluated the average coverage rate, a useful metric in the case of PE38 given its large number of epitopes, as this metric more directly evaluates whether each variant is able to cover multiple epitopes on its own. As shown in Fig. 4.6 left, on average our top designs hit 50% (14 epitope residues from 7 epitope clusters) of the identified epitope residues for mutation loads of at least 8, and the coverage rates increase linearly with the mutation loads. Given the large number of epitopes in PE38 (e.g., as compared to the single epitope we were targeting in EGFP), high recall is important, and thus in this case a relatively low probability threshold tended to yield better coverage rate. Union coverage rate is naturally higher; e.g., the union of the top three 6-mutation designs covers more than 80% of the epitope residues (Fig. 4.6 right). While high average coverage implies that each design own its own hits a good number of epitopes, high union coverage indicates that they do so in diverse ways; i.e., if multiple designs were selected, they would not simply be duplicates.

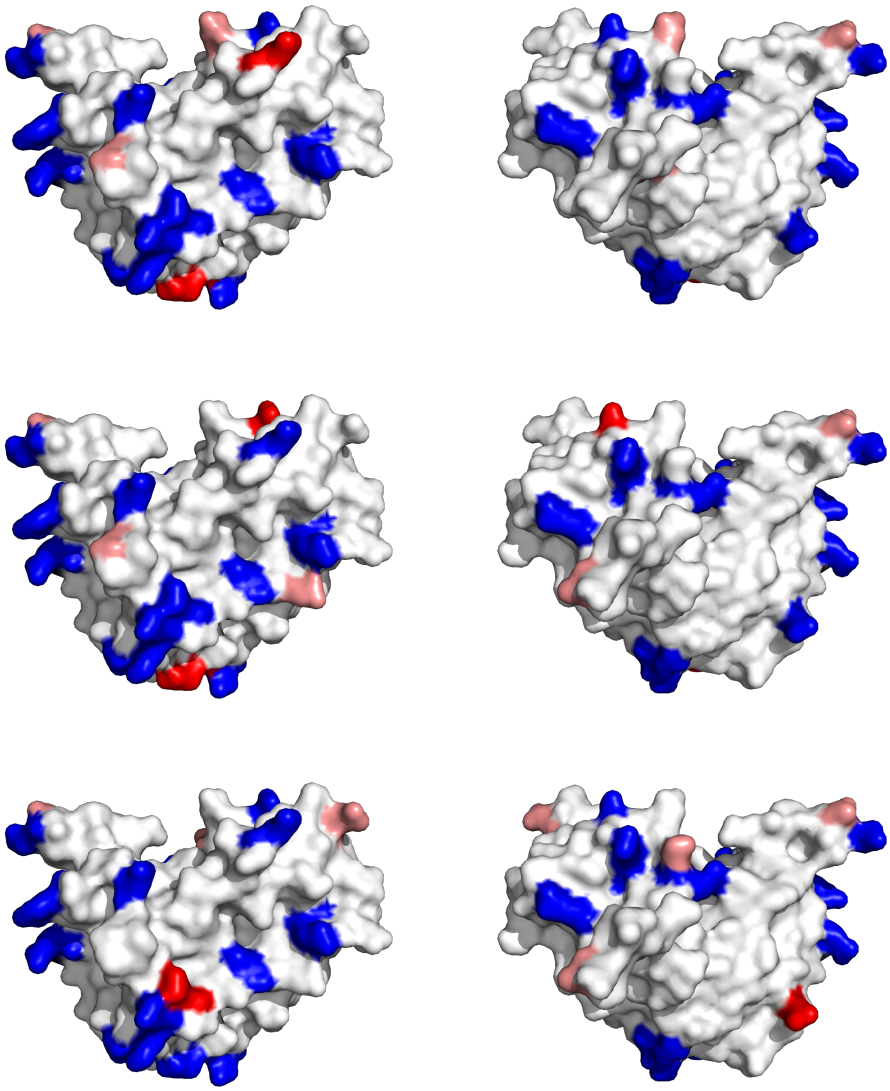

Figure 4.5: Front (left) and back (right) of the top three 6-mutation PE38 B cell epitope deletion designs. Residues highlighted on the structure show mutations hitting known epitopes (red), other known but not directly mutated epitope residues (blue), and mutations not directly hitting (but generally covering) known epitopes (salmon).

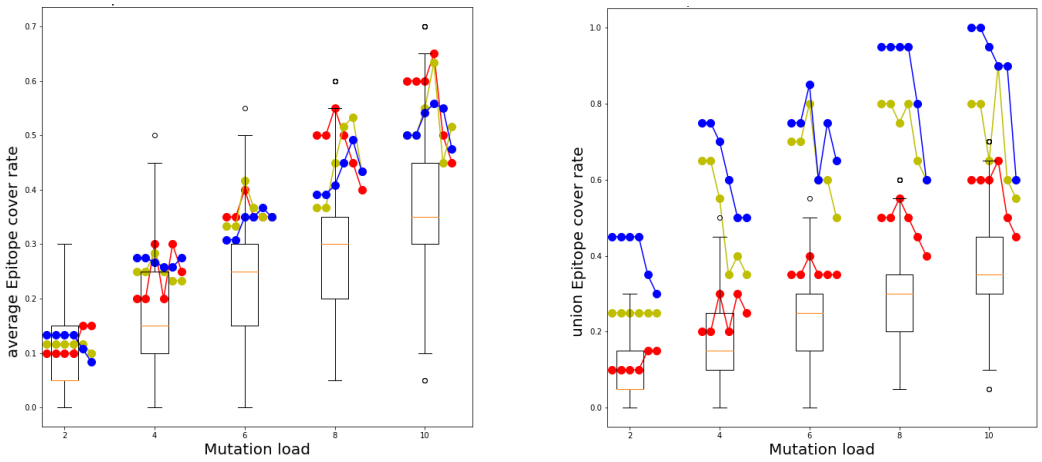

Figure 4.6: Average (left) and union (right) coverage of known PE38 epitope residues by designs (lines) and random sets of mutations (boxplots) at different mutation loads (groups on the x axis). Designs are chosen at epitope prediction probability thresholds (points on the lines from 0.0 to 0.5) and the coverage over the top several designs (red: 1, yellow: 3, blue: 6) is used to characterize design quality and diversity.

Benchmark. To investigate the expected performance of our algorithm more broadly, we applied it to the same set of 41 Ab-Ag complexes selected from SAbDab, applying PInet to predict epitopes but eliminating the binding partner from the representative Ab set. We kept a low mutation load, from 1 to 4, which yields designs more likely to be stable and practice, and was sufficient to display strong trends since for these there was only one (unknown) Ab each. We took the union of either the top three or top six designs (thereby assessing both the quality and the diversity of the coverage provided by the designs). Random sets of mutations of the same size, i.e., summed over the designs, so that for example the top 3 designs at mutation load 2 would be compared to 6 random mutations. Fig. 4.7 shows that our sampling algorithm always outperforms random in terms of median epitope coverage rate. The difference increases with mutation load, as would be expected for our sampling process, which seeks to spread mutations over the surface.

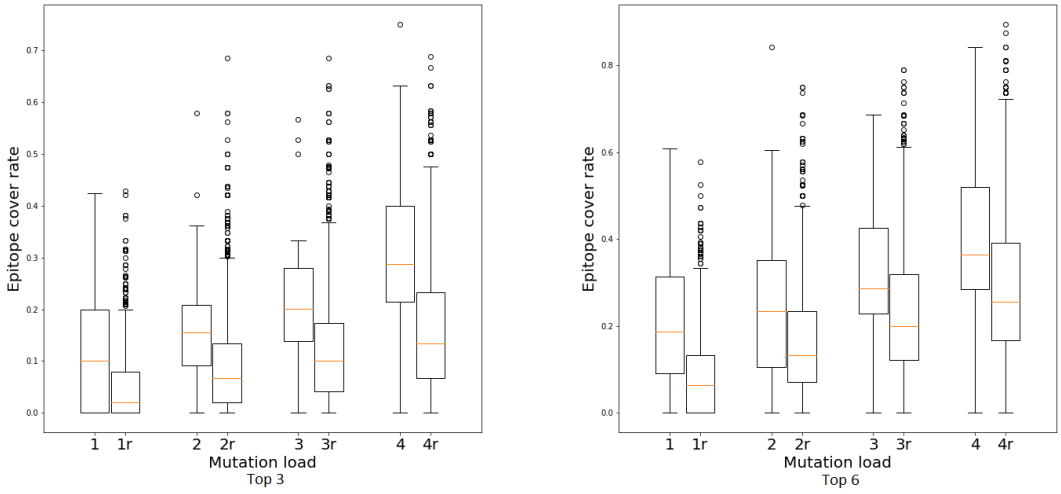

Figure 4.7: Coverage of epitopes in a 41 Ab-Ag dataset by top designs (left of each pair of boxes) or a corresponding number of random mutations (right of each pair), for either the top three (left panel) or the top six (right panel) designs.

4.2.3. Epitope Map Deconvolution

Hen egg white lysozyme (HEL) serves as a good test case for epitope map deconvolution, due to the availability of a number of structures with HEL in complex with Abs of different epitopes [149]; we here consider four different complexes (PDB ids 1BVK, 1MLC, 1DQJ, and 2I25). In order to build an epitope map, each epitope must be disrupted at least once; thus a key goal, and our main quality metric for designs, is high epitope coverage rate. The number of mutations per variant, i.e., per epitope region, is controlled by the patch size, with a larger patch size leading to more mutations per variant disrupting more co-located epitopes and a smaller patch size using fewer mutations each. The smaller patch size may be safer from the perspective of likelihood of maintaining Ag stability, but will require more variants in the panel to achieve sufficient mapping. With these aspects in mind, we quantified the epitope coverage of the top panels, over different patch sizes and epitope probability thresholds. We again compared to random designs as a baseline, here using the same number of mutation as the total number of mutations within the corresponding panel.

Fig. 4.8 shows that our designs generally cover 80-90% of epitope residues, much better than random sets of mutations of corresponding sizes. As discussed, smaller patch sizes (8) and lower prediction probability thresholds (0.1) lead the panel to need more variants. As we saw with PE38, lower epitope probability thresholds lead to better coverage, but even at higher thresholds the coverage is still more than 80% for patch sizes larger than 8, and the panels are smaller. At the larger patch sizes, the panels can cover 95% of the epitope residues with fewer than 20 variants per panel.

The goal of epitope map deconvolution is, upon obtaining binding data for the Abs against the Ag variant panel, to be able to assign different Abs to different epitopes on the Ag. To gain insights into how well our panels might be able to do that, we further explored the differences in expected binding disruption of the anti-

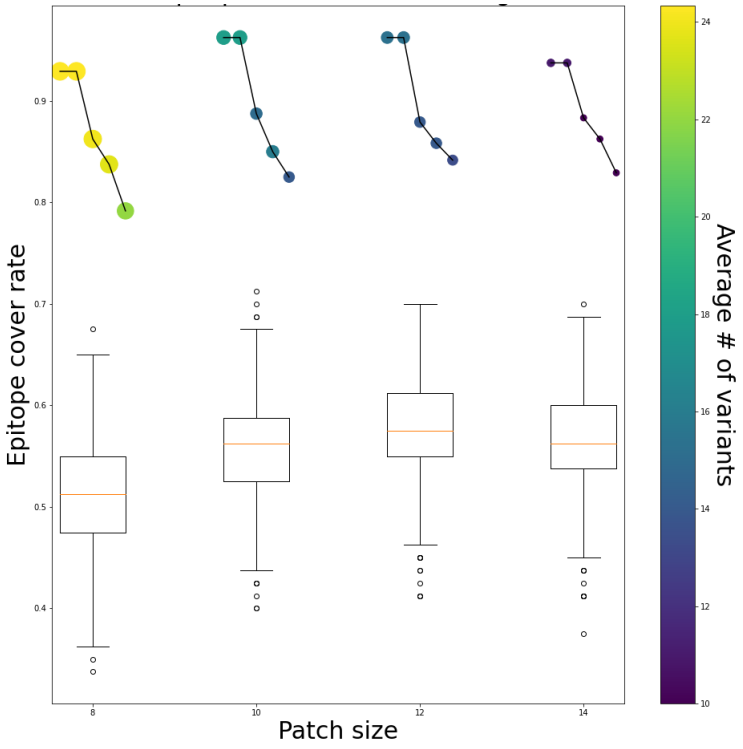

Figure 4.8: HEL epitope map deconvolution, evaluating the average epitope coverage by the top three designs while varying patch size (groups on the x axis), epitope probability (points on lines, ranging from 0.0 on the left to 0.4 on the right). Lighter color and larger size marker for the points indicate a larger panel. Boxplots are randomly sampled designs with the same number of mutations as an average panel at the given patch size.

HEL Abs with respect to a designed HEL panel (patch size 10, epitope probability threshold 0.5). Fig. 4.9 shows Ab "fingerprints" according to the panel, defined such that if a mutation of a variant covers the Ag's epitope region (within 12Å of any epitope residue), it will disrupt the binding. We recognize that one mutation may not be sufficient (though that's the basis for alanine scanning), and the mutation may destabilize the Ag, but this fingerprint nonetheless paints a picture of, in a good scenario, how the different variants could be expected to deconvolve the different Abs. The fingerprints in Fig. 4.9 make clear that these four Abs display quite different expected disruption profiles against the designed panel of 18 variants, and that most of

the variants help distinguish different epitopes. There are some obvious redundancies (e.g., 1 and 2, 4 and 5, etc.), which might lead to robustness, or might indicate wasted experimental effort. Suppl. Fig. 4.13 further evaluates the fraction of variants in each panel that are expected to uniquely contribute to the deconvolution process (55% in this case). We also see that by reducing the design space with PInet, we are able to increase the fraction of variants contributing to deconvolution.

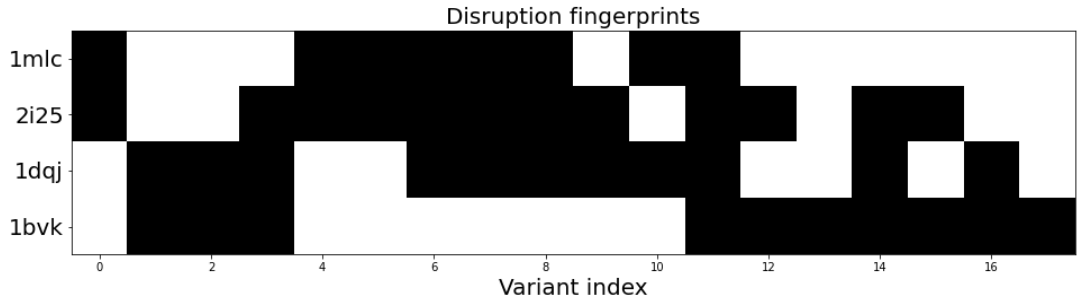

Figure 4.9: Visualization of Ab fingerprints for anti-HEL Abs (PDB ids 1BVK, 1MLC, 1DQJ, and 2I25), rows, against a designed panel of 18 variants, columns. A black cell indicates that the variant covers an Ab epitope and thus may be expected to provide a disruption signal in a good scenario.

Section 4.3

Conclusion

We developed algorithms to design Ag variants that escape recognition by an entire set of Abs, even in the case where those Abs are uncharacterized. This ability can enable the generation of variants of a therapeutic protein that escape a preexisting polyclonal response, as well as the development of a panel of variants by which to separate a polyclonal mixture of Abs by their specificities while also localizing them to their epitopes on their Ag. By leveraging the information regarding Ab-Ag specificity encoded in a cutting-edge epitope predictor, we focus design effort on the broadly most important predicted epitopes on the Ag surface, according to a set of representative

Abs that sufficiently approximate the paratope diversity exhibited by the true but unknown Abs. Our geometry-based sampling algorithms then optimize either single variants for B cell epitope deletion or a complementary panel of variants for epitope map deconvolution, selecting sites and amino acid substitutions predicted to cover the important predicted epitope residues, balancing the goals of reducing Ab binding while also maintaining Ag stability.

Strikingly, our benchmark evaluation showed that epitope prediction performance based on a suitably large representative set of Abs was nearly as good as that with the true Ab, and the performance remained overall quite good, though more variable, even with only 10 representative Abs. This suggests that the paratope diversity in the representative set is good enough to capture the true paratope, to within the "resolution" of the epitope predictor employed here. As epitope predictors improve, or if the design method were extended to leverage finer grained information about the interaction (e.g., pairs of contacting residues), the representative Ab approach may need to be elaborated. Using models of germline Abs may better capture the diversity of paratopes in the true polyclonal mixture, as the polyclonal Abs would have been matured from such germline Abs while maintaining the same general binding preferences. Alternatively, models of large Ab repertoires would represent a greater diversity of high-resolution paratope variation, and suitable clustering and weighting among their predictions may capture preferences of a particular polyclonal response.

Our Ag variant design is driven by a geometry-based sampling algorithm focused on covering the residue positions with highest predicted epitope probability over the representative Abs. The algorithms subsequently chooses amino acid substitutions at those positions that are likely to disrupt Ab binding but not Ag stability, according to predictive models. We currently generate a number of different designs and then rank them by predicted epitope coverage and Ag stability. An extension of the algorithm

could integrate these objectives, allowing the site selection to directly account for the predicted effects on Ab binding and stability, in addition to coverage of predicted epitopes. We showed that post-processing the designs for diversity can yield better epitope coverage over the subset of tested variants; an extended algorithm could optimize for diversity of a set of variants during the design process. While the evaluation of binding disruption employed here does not account for predicted details of the Ab-Ag interaction, with improving prediction capabilities (and, as discussed above, a suitably extended representative Ab set), those details could be leveraged to more precisely target disruptive residue interactions. More generally, our method can readily incorporate other metrics regarding the effects of mutations on Ag stability and Ab-Ag interaction, accounting simultaneously for the combined effects of mutations on both properties. Our results demonstrate that it is possible to design Ag variants to escape recognition by sets of uncharacterized Abs, and our methods provide a strong basis for future developments in such wide-ranging Ab escape applications.

Section 4.4

Methods

As illustrated in Fig. 4.1, our algorithms for B cell epitope deletion and epitope map deconvolution design either single variants of a given Ag (B cell epitope deletion) or panels of variants of a given Ag (epitope map deconvolution) based on the predicted effects of Ag mutations on both Ag stability and predicted recognition by existing Abs. We detail here the general case where no information is available regarding the Abs or their epitopes; it is straightforward to utilize any such information to focus the design accordingly.

4.4.1. Preprocessing

The preprocessing for design starts with a target Ag structure and a set of representative Ab structures, predicts putative epitopes on the Ag for the Abs, prefilters allowed mutations to the Ag, and predicts the effects of mutations on Ab binding and Ag stability.

Target Ag. We are employing a structure-based approach, and thus require either the Ag's structure from the PDB [153] or a high-quality model based on its sequence [78, 77, 6].

Representative Abs. While we don't have any specific information about the actual Abs that recognize the target Ag, we hypothesize that, due to their common structure and origins (i.e., derivation from germline-based progenitors), a diverse set of *representative Abs*, while not actually specific to the target, represent the paratopes of the actual Abs to a sufficient approximation for use in epitope prediction. Here we take as representative Abs a set of high quality (max resolution 3.0Å), low sequence similarity (max sequence identity 60%) human and murine Abs filtered from the curated antibody structure database SabDab [42].

Epitopes. In the absence of information regarding epitope localization, epitopes are predicted for the representative Abs, with the aim of including many/most of their true epitopes (at the cost of accepting false positives as well). We assume that the epitope prediction method (here we use PINet [33]) labels each residue with a probability of being an epitope residue, and we take as P_e the max probability over all representative Abs predictions.

Allowed mutations. In principle, any mutation to the Ag could be considered; in practice, it is often helpful to prefilter to a set M of allowed mutations that are likely

to maintain the stability and functionality of the target protein while still being able to disrupt binding [165, 21]. Here we use a method based on amino acid frequency (more than 35%) among a set of homologs of the target filtered for quality and diversity [28]. Furthermore, since we are seeking to disrupt Ab binding, we only consider positions that are solvent accessible with relative solvent accessibility (RSA) larger than 0.15; while other residues may indirectly affect Ab recognition, such surface residues are more likely to have the desired effect.

Disruptive mutations. Since we do not have Ab-Ag complex structures, we use sequence-based methods to evaluate the potential disruptiveness of an Ag mutation on Ab binding. (Note that the method could readily be extended to incorporate evaluation of disruptiveness based on docking models, if docking is used to predict epitopes [98].) Intuitively, changing physicochemical properties such as size and charge is likely to disrupt specific Ag binding. To capture this intuition, we use BLOSUM62 [64], which, though derived from and for sequence alignment, generally captures commonality of amino acid types. Other matrices such as PAM [35] or EDSSMat [157] could readily be incorporated. We filter the set of allowed mutations M to a subset M_d that are sufficiently disruptive, in that their BLOSUM scores are low enough but not too low, namely -2, -1, 0, or 1. Since the BLOSUM score only provides a sense of similarity between amino acid types, we only use it as a filter for disruptive mutations to consider, as opposed to a quantitative property to optimize.

Coverage score. A mutation changes the shape and characteristics not just of the amino acid side chain, but also of the general surface around it. Thus, we say that a mutation *covers* an epitope residue if that residue is sufficiently close (less than 12Å between centroids). For a set of mutations (i.e., an Ag variant), we then compute the coverage score as the sum of the predicted probability P_e for all epitope residues

which are covered by at least one mutation. Coverage scores here thus represent the expected value of the number of epitope residues covered by the set of mutations.

Stability score. A mutation may also affect the stability and function of the Ag. To evaluate these effects, we compute the probabilities P_s for each possible mutation, according to the protein language model ProtTrans [44], noting that other such models such as ESM [135] or even structure-based methods methods such as Rosetta [3] could readily be used. For the results presented here, we follow the approach of treating mutations as conditionally independent given the wild-type sequence [101], and take the sum of the mutation scores as the measure of stability for a multiple-mutation variant. The fact that the mutations are at solvent exposed sites, to residues more likely to interact with the Ab than with each other, lends some support to this assumption. In the future, our method could use other scoring approaches to properly account for the combinatorial interactions required to compute correct conditional probabilities.

4.4.2. B cell epitope deletion

The general goal of our B cell epitope deletion algorithm is to generate and rank a set of stable and functional variants of an Ag that are predicted to evade recognition by (unknown) preexisting Abs. More precisely, based on the preprocessing above, we are given a target protein with residues R, predicted epitopes $E \subset R$ for representative Abs, and allowed disruptive mutations M_d. We then generate a set of variants of fixed mutation load l (this can be varied for different design runs), ranked by coverage and stability scores.

To optimize coverage, we perform Farthest Point Sampling (FPS) on a graph representation of the surface residues of the Ag. In particular, the graph has nodes for the surface residues and edges connecting residues that are close enough; here we

use a 10$\mathring{A}$ threshold on the geodesic distances between residue centroids [126]. FPS is applied to those residues predicted to be epitopes, i.e., with P_e meeting a threshold (we evaluate the effects of the threshold from 0 to 0.5). FPS samples points iteratively, at each step selecting the point that is geodesically furthest from the set so far. The selection is repeated until reaching the specified mutation load. Given the selected positions, we select an amino acid substitution for each from the allowed mutations at positions within a small area (residues which are within the coverage radius, 12$\mathring{A}$ between centroids of selected positions), according to the highest stability score.

Since the output of FPS depends on the first point being sampled, we repeat the process, using each of the predicted epitope residues as the starting point. This thereby generates a number of different variants, which we then rank by coverage and stability scores. In particular, for variant i defined by mutations $M_i \subseteq M_d$ covering residues $R_i \subseteq R$, the combined score S_i is:

$$S_i = a_e \eta \left(\sum_{r \in R_i} P_e(r) \right) + a_m \eta \left(\sum_{m \in M_i} P_s(m) \right) \tag{4.1}$$

where η normalizes a given score based on the corresponding mean and standard deviation of all sampled designs and a_e, a_m weight the two scores (results shown use a weight of 1 for each). The normalization puts these two different scores on common footing; we note that a Pareto optimization approach [63] could be taken instead of fixing weights.

4.4.3. Epitope map deconvolution

While B cell epitope deletion seeks a single variant that can escape existing Abs, in contrast epitope map deconvolution seeks a panel of variants that together reveal the epitopes of existing Abs. Each variant only needs to escape Abs localized to a particular epitope region, but together the panel should cover all epitopes (and

as before, the variants still need to be stable). More precisely, we are given a target protein with residues R, predicted epitopes $E \subset R$ for representative Abs, and allowed disruptive mutations M_d, we generate a set of variants with maximal panel size p and maximal mutational load per variant v (both can be varied in different design runs).

Since each variant in a deconvolution panel targets one particular (spatially localized) epitope region, the design algorithm is different from B cell epitope deletion. We again follow a greedy sampling approach, but rather than taking the farthest point next, we instead take the closest point next, as long as it's not "too close" (at least a given radius r geodesically, we here choose r to be 10Å) and thus defines a different epitope region than those sampled so far. As a result, each point defines the centroid of a region, and a set of sampled points is chosen so as to cover the predicted epitopes by surface patches around these points. The sampling continues until everything is covered or we reach the limit p on the panel size. The radius r could be adjusted manually according to p and v to control the overlapping between patches so as to control the resolution of the epitope map. Given the sampled points / patches, we then design variants to disrupt each one. For this, we use the same FPS algorithm as descried, but now limited to the specified region, thereby covering the particular epitope region with disruptive mutations.

Again, different initial starting points yield different designs, and we rank them based on overall stability and coverage. Consider a panel P_j, i.e., set of variants $V_1, \ldots, V_p$, where variant V_i covers epitope residues R_i using mutations M_i. The score of panel P_j is then:

$$S_j = a_e \eta \left(\sum_{V_i \in P_j} \sum_{r \in R_i} P_e(r) \right) + a_k \eta \left(\sum_{V_i \in P_j} \sum_{m \in M_i} P_s(m) \right) \tag{4.2}$$

where again η normalizes stability and coverage scores by all design scores, and a_e, a_m weight the two components; we used 1.0 but different weightings or a Pareto opti-

mization approach could be employed. Notice that while the stability score is the same for both epitope deletion and epitope map deconvolution design, the coverage score is different. In particular, for deconvolution design, a residue can be covered by different variants in the designed panel, and thus can contribute multiple times to the coverage score. Summing all a residue's contributions accounts for some overlap between patches, which may help provide additional deconvolution information, e.g., to help pinpoint whether an epitope is at the boundary or center of a patch.

Section 4.5

Supplementary Figures

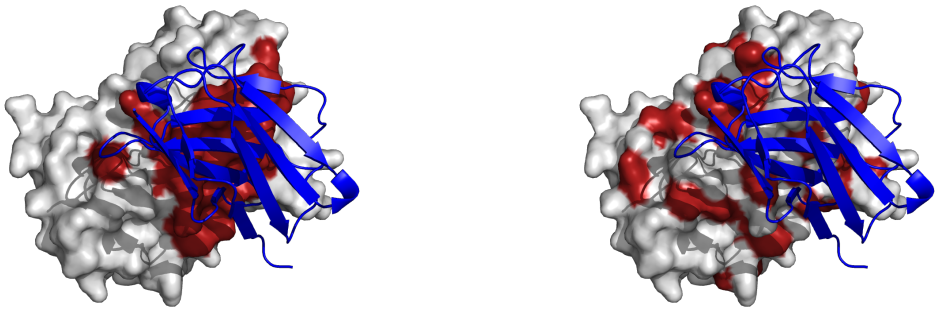

Figure 4.10: EGFP (surface) and nanobody (cartoon) complex structure (PDB id 3OGO), with epitope residues highlighted red. (Left) true epitopes; (Right) predicted epitopes by PInet using representative Abs.

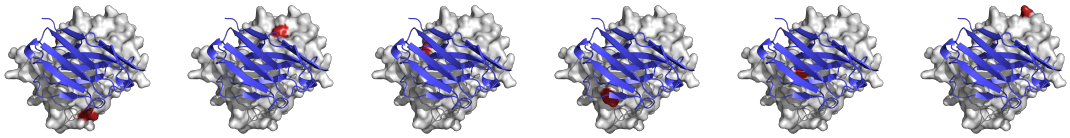

Figure 4.11: Top six 6-mutation (bottom row) diversity-filtered epitope-deleting EGFP designs (red: mutated sites; blue cartoon: nanobody, for illustration but not used in design).

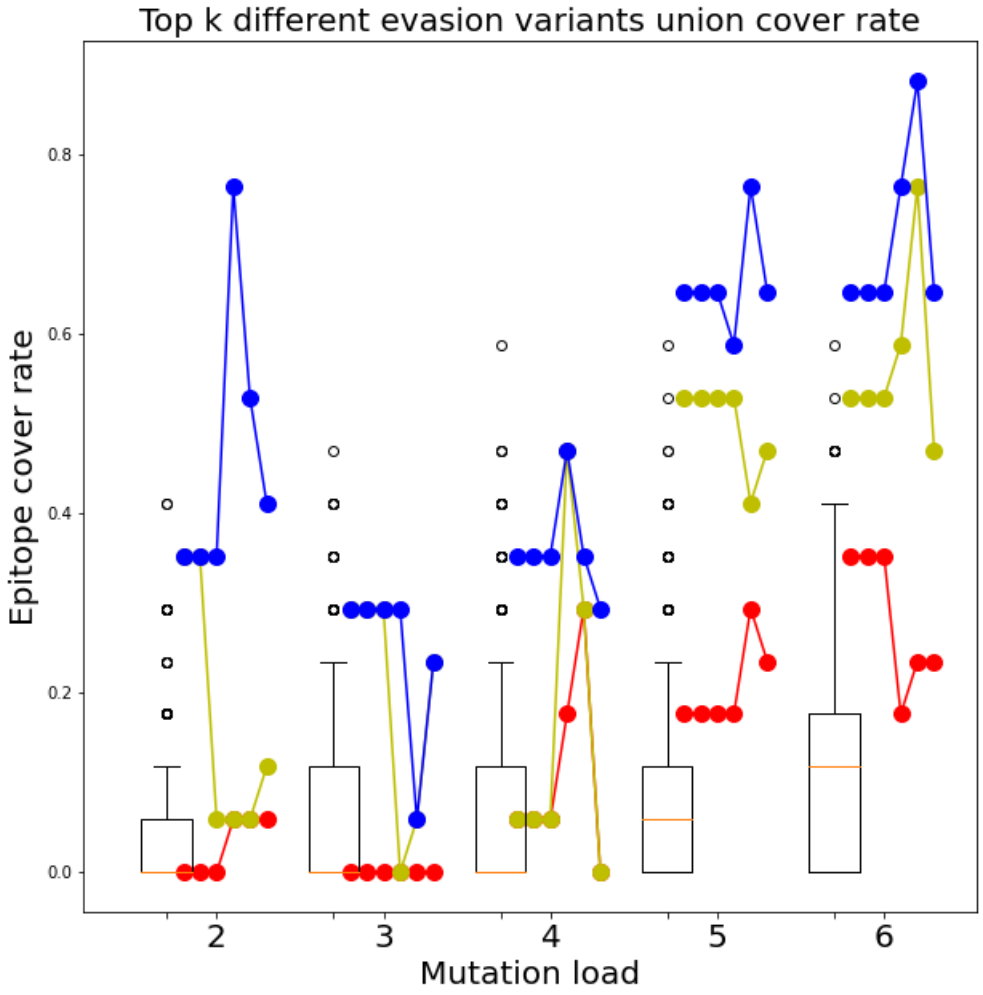

Figure 4.12: Coverage of known EGFP epitope residues by diversity-filtered designs (lines) and random sets of mutations (boxplots) at different mutation loads (groups on the x axis). Designs are chosen at epitope prediction probability thresholds (points on the lines from 0.0 to 0.5) and the coverage over the top several sufficiently different designs (red: 1, yellow: 3, blue: 6) is used to characterize design diversity.

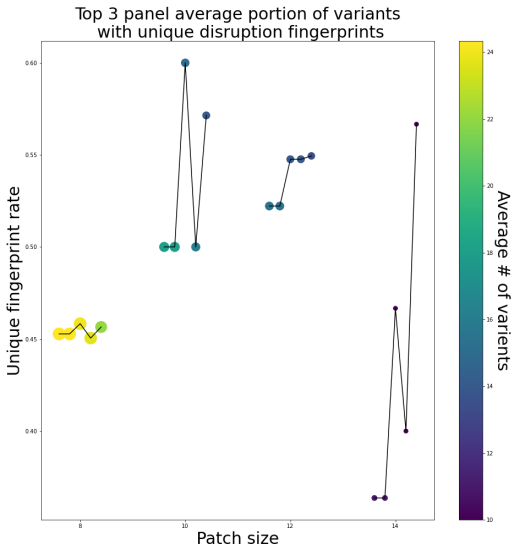

Figure 4.13: Fraction of variants in the anti-HEL panel that are expected to contribute to epitope map deconvolution (i.e., have a unique fingerprint).

Chapter 5

Discussion

Due to the importance of protein interactions in wide-ranging cellular processes, methods for experimentally characterizing these interactions have been pursued for decades. Experimental limitations, including capacity, time, and expense, have led to significant interest in the development of computational methods to predict interactions. Due to the relative sparsity of experimental data, for many years the most successful methods were typically based primarily on physical modeling. With experimental advances, more data began to accumulate, and traditional machine learning methods began to move to the forefront for some problems. We have now reached a point where there is sufficient data, and the modeling approaches have sufficiently advanced, to enable deep learning methods to flourish, particularly when the still relatively limited input data can be augmented, e.g., by leveraging physical modeling or by performing unsupervised pre-training with richer datasets. This thesis demonstrates the utility of such deep learning approaches by developing and applying innovative methods that combine data-driven and physical modeling in three different protein recognition applications: predicting protein-protein interactions, predicting protein-glycan interactions, and designing proteins to escape interactions. Together, these contributions not only provide the community with better predictive and de-

sign tools, but also demonstrate the ability of advanced modeling methods to extract useful information from limited and noisy interaction data.

In predicting protein-protein interactions, PInet improved the state-of-the-art by leveraging the geometry of protein surfaces, which mediate interactions, and employing geometric deep learning to represent the underlying contributions to specific recognition. More generally, the results from this study imply that such geometry-based models and algorithms could provide a general framework for other PPI-related tasks including both finer-resolution tasks such as predicting binding affinities to enable affinity maturation, as well as coarser-resolution tasks such as predicting whether or not a pair of proteins are cognates to enable virtual screening. Major drawbacks of the current models and methods include not directly dealing with the (arbitrary) initial orientations of the partners, and not accounting for protein flexibility and conformational change upon binding. Recent research has demonstrated that the orientation problem is solvable via an appropriate encoding of the structure (e.g. SE(3)-equivariant). However, suitably handling conformational change remain very challenging, with potential approaches (beyond simply sampling multiple conformations) including employing probabilistic geometric representations, or combining geometric and sequence models in multimodal learning in the hopes that the structural and sequence representations encode complementary information regarding possible changes. With rapidly continuing advances in Computer Vision (geometry), Natural Language Processing (sequence), and Multimodal Learning, I am confident that better interaction models will be able to leverage these to better model and predict the protein-protein interactions.

In predicting protein-glycan interactions, GlyBERT demonstrated that training a well-structured embedding in a suitable latent space establishes a strong foundation that directly enables improvements in downstream tasks via fine tuning. In some con-

texts, such as this one, clean and well labeled data are limited; GlyBERT overcomes this by using high quality input features and unsupervised pre-training. We showed that, upon fine-tuning, the model performed well on a number of different tasks, and hope that our well-pre-trained base model may likewise benefit the larger community after they fine-tune it with other data for other tasks. We also demonstrated the ability of GlyBERT to support molecule optimization or generation, in a case study of modifying a glycan's predicted immunogenicity. Though classical methods such as Markov chain Monte Carlo (MCMC) or simulated annealing (SA) could do such design with any model, the gradient-based approach enabled by GlyBERT is faster and also allows better control of multiple objectives. While GlyBERT relies on a relatively small amount of data for unsupervised pre-training, glycobiologists are pursuing new shotgun and microarray methods that, while of potentially lower fidelity than the assays used here, are higher-throughput and thus may provide a richer basis for pre-training. Furthermore, GlyBERT was based on an encoding of the glycan alone, while predictions could benefit from a joint representation of the glycan and a protein that has recognized it or to which it has been attached (e.g., glycosylation or glycation). Developing high-quality datasets and models for such joint modeling should ba e fruitful avenue to improve prediction capabilities. Finally, while the field has mostly focused on classifying glycans, we showed here that it is possible to use predictive models "in reverse", designing modifications to glycans, e.g., to change a property such as immunogenicity. Experimental validation of such predictions, or even computational-experimental iterative exploration of glycan space, could both improve models and also discover beneficial new glycans and interactions.

In designing proteins to evade Ab recognition, we showed that a deep learning model of Ab recognition captures the general properties of epitope specificity well enough to enable representative Abs to stand in for unknown true Abs and enable us to

design proteins so as to evade existing Abs. We developed geometry-based algorithms to suitably sample epitope residues and regions in order to delete B cell epitopes and to design panels for epitope map deconvolution. Critical to the approach are the set of representative Abs and the ability to use them to predict the epitopes of the unknown Abs. Methods for epitope prediction are continuing to improve, and the method could potentially move from using just a prediction of the Ag epitope residues to a prediction of the interacting Ab paratope : Ag epitope residue pairs. Furthermore, the representative Abs could possibly be replaced by models of repertoire Abs identified via deep sequencing. B cell epitopes are usually ignored during early stage drug discovery due to the complexity of dealing with them, a promising B cell deletion algorithm could help with early stage candidate filtering as well as late stage lead optimization. Epitope map deconvolution and related computational-experimental methods for mapping and localizing epitopes are relatively new concepts, but hold much promise to guide the selection and development of therapeutic Abs, vaccine Ags, and so forth.

Ultimately, the lessons learned from these complementary studies point to general next steps in improving computationally-driven protein design. One key is to learn a suitable latent space in which to represent the targets of design (proteins, glycans, interacting partners, ...), integrating available information from both sequence and structure. On its own, this will improve prediction of epitopes, disruption, stability, glycan-binding, etc. It will also enable design processes to take advantage of suitable representations in optimizing modifications. While our design algorithms employed greedy sampling, I envision that protein design will move toward continuously sampling trajectories with multiple objectives and constraints applied at the same time. Gradients (or energy scores for MCMC and SA) from different objectives will be computed with respect to the embedding in the protein latent space and point the

directions for improving the molecule in order to generate better variants. Experimental data collected from such designs could then be fed back to improve the models and the generated molecules. Such a framework promises to reduce protein engineering cycle times while also improving hit rates, ultimately leading to better leads for further development.

Bibliography

[1] W Mark Abbott, Melissa M Damschroder, and David C Lowe, *Current approaches to fine mapping of antigen–antibody interactions*, Immunology **142** (2014), no. 4, 526–535.

[2] Fayyaz ul Amir Afsar Minhas, Brian J Geiss, and Asa Ben-Hur, *Pairpred: Partner-specific prediction of interacting residues from sequence and structure*, Proteins: Structure, Function, and Bioinformatics **82** (2014), no. 7, 1142–1155.

[3] Rebecca F Alford, Andrew Leaver-Fay, Jeliazko R Jeliazkov, Matthew J O'Meara, Frank P DiMaio, Hahnbeom Park, Maxim V Shapovalov, P Douglas Renfrew, Vikram K Mulligan, Kalli Kappel, et al., *The rosetta all-atom energy function for macromolecular modeling and design*, Journal of chemical theory and computation **13** (2017), no. 6, 3031–3048.

[4] Juan-Manuel Anaya, Yehuda Shoenfeld, Adriana Rojas-Villarraga, Roger A Levy, and Ricard Cervera, *Autoimmunity: from bench to bedside [internet]*, (2013).

[5] anytree, *anytree version 2.8.0*, `https://anytree.readthedocs.io/en/2.8.0/`, 2020.

[6] Minkyung Baek, Frank DiMaio, Ivan Anishchenko, Justas Dauparas, Sergey Ovchinnikov, Gyu Rie Lee, Jue Wang, Qian Cong, Lisa N Kinch, R Dustin

Schaeffer, et al., *Accurate prediction of protein structures and interactions using a three-track neural network*, Science **373** (2021), no. 6557, 871–876.

[7] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio, *Neural machine translation by jointly learning to align and translate*, arXiv preprint arXiv:1409.0473 (2014).

[8] Nathan A Baker, David Sept, Simpson Joseph, Michael J Holst, and J Andrew McCammon, *Electrostatics of nanosystems: application to microtubules and the ribosome*, Proceedings of the National Academy of Sciences **98** (2001), no. 18, 10037–10041.

[9] Xu Benjin and Liu Ling, *Developments, applications, and prospects of cryo-electron microscopy*, Protein Science **29** (2020), no. 4, 872–882.

[10] Helen M Berman, Tammy Battistuz, Talapady N Bhat, Wolfgang F Bluhm, Philip E Bourne, Kyle Burkhardt, Zukang Feng, Gary L Gilliland, Lisa Iype, Shri Jain, et al., *The protein data bank*, Acta Crystallographica Section D: Biological Crystallography **58** (2002), no. 6, 899–907.

[11] Ola Blixt, Steve Head, Tony Mondala, Christopher Scanlan, Margaret E Huflejt, Richard Alvarez, Marian C Bryan, Fabio Fazio, Daniel Calarese, James Stevens, et al., *Printed covalent glycan array for ligand profiling of diverse glycan binding proteins*, Proceedings of the National Academy of Sciences **101** (2004), no. 49, 17033–17038.

[12] Daniel Bojar, Diogo M Camacho, and James J Collins, *Using natural language processing to learn the grammar of glycans*, bioRxiv (2020).

[13] Daniel Bojar, Rani K Powers, Diogo M Camacho, and James J Collins, *Sweet-origins: Extracting evolutionary information from glycans*, bioRxiv (2020).

[14] ———, *Deep-learning resources for studying glycan-mediated host-microbe interactions*, Cell Host & Microbe **29** (2021), no. 1, 132–144.

[15] Thomas Bourquard, Astrid Musnier, Vincent Puard, Shifa Tahir, Mohammed Akli Ayoub, Yann Jullian, Thomas Boulo, Nathalie Gallay, Hervé Watier, Gilles Bruneau, et al., *Mabtope: a method for improved epitope mapping*, The Journal of Immunology **201** (2018), no. 10, 3096–3105.

[16] Bryan Briney, Anne Inderbitzin, Collin Joyce, and Dennis R Burton, *Commonality despite exceptional diversity in the baseline human antibody repertoire*, Nature **566** (2019), no. 7744, 393.

[17] Benjamin D Brooks, Adam Closmore, Juechen Yang, Michael Holland, Tina Cairns, Gary H Cohen, and Chris Bailey-Kellogg, *Characterizing epitope binding regions of entire antibody panels by combining experimental and computational analysis of antibody: antigen binding competition*, Molecules **25** (2020), no. 16, 3659.

[18] Benjamin D Brooks, Adam R Miles, and Yasmina N Abdiche, *High-throughput epitope binning of therapeutic monoclonal antibodies: why you need to bin the fridge*, Drug discovery today **19** (2014), no. 8, 1040–1044.

[19] Jason Cory Brunson, *Ggalluvial: layered grammar for alluvial plots*, Journal of Open Source Software **5** (2020), no. 49, 2017.

[20] Rebekka Burkholz, John Quackenbush, and Daniel Bojar, *Using graph convolutional neural networks to learn a representation for glycans*, Cell Reports **35** (2021), no. 11, 109251.

[21] Oliver Buß, Jens Rudat, and Katrin Ochsenreither, *Foldx as protein engineering tool: better than random based approaches?*, Computational and structural biotechnology journal **16** (2018), 25–33.

[22] Eric J. Carpenter, Shaurya Seth, Noel Yue, Russell Greiner, and Ratmir Derda, *Glynet: A multi-task neural network for predicting protein-glycan interactions*, bioRxiv (2021).

[23] Rong Chen, Li Li, and Zhiping Weng, *Zdock: an initial-stage protein-docking algorithm*, Proteins: Structure, Function, and Bioinformatics **52** (2003), no. 1, 80–87.

[24] Tianqi Chen and Carlos Guestrin, *Xgboost: A scalable tree boosting system*, Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, 2016, pp. 785–794.

[25] Hao D Cheng, Sebastian K Grimm, Morgan SA Gilman, Luc Christian Gwom, Devin Sok, Christopher Sundling, Gina Donofrio, Gunilla B Karlsson Hedestam, Mattia Bonsignori, Barton F Haynes, et al., *Fine epitope signature of antibody neutralization breadth at the hiv-1 envelope cd4-binding site*, JCI insight **3** (2018), no. 5.

[26] Kai Cheng, Yusen Zhou, and Sriram Neelamegham, *Drawglycan-snfg: a robust tool to render glycans and glycopeptides with fragmentation information*, Glycobiology **27** (2017), no. 3, 200–205.

[27] Yoonjoo Choi, Jacob M Furlon, Ryan B Amos, Karl E Griswold, and Chris Bailey-Kellogg, *Disruppi: structure-based computational redesign algorithm for protein binding disruption*, Bioinformatics **34** (2018), no. 13, i245–i253.

[28] Yoonjoo Choi, Deeptak Verma, Karl E Griswold, and Chris Bailey-Kellogg, *Episweep: Computationally driven reengineering of therapeutic proteins to reduce immunogenicity while maintaining function*, Computational Protein Design, Springer, 2017, pp. 375–398.

[29] Lachlan Coff, Jeffrey Chan, Paul A Ramsland, and Andrew J Guy, *Identifying glycan motifs using a novel subtree mining approach*, BMC bioinformatics **21** (2020), no. 1, 42.

[30] Stephen R Comeau, David W Gatchell, Sandor Vajda, and Carlos J Camacho, *Cluspro: an automated docking and discrimination method for the prediction of protein complexes*, Bioinformatics **20** (2004), no. 1, 45–50.

[31] Richard D Cummings, *The repertoire of glycan determinants in the human glycome*, Molecular BioSystems **5** (2009), no. 10, 1087–1104.

[32] Brian C Cunningham and James A Wells, *High-resolution epitope mapping of hgh-receptor interactions by alanine-scanning mutagenesis*, Science **244** (1989), no. 4908, 1081–1085.

[33] Bowen Dai and Chris Bailey-Kellogg, *Protein interaction interface region prediction by geometric deep learning*, Bioinformatics (2021).

[34] Bowen Dai, Daniel E Mattox, and Chris Bailey-Kellogg, *Attention please: modeling global and local context in glycan structure-function relationships*, bioRxiv (2021).

[35] M Dayhoff, R Schwartz, and B Orcutt, *22 a model of evolutionary change in proteins*, Atlas of protein sequence and structure **5** (1978), 345–352.

[36] Miranda de Graaf and Ron AM Fouchier, *Role of receptor binding specificity in influenza a virus transmission and pathogenesis*, The EMBO journal **33** (2014), no. 8, 823–841.

[37] Sjoerd J De Vries, Aalt DJ Van Dijk, Mickaël Krzeminski, Mark van Dijk, Aurelien Thureau, Victor Hsu, Tsjerk Wassenaar, and Alexandre MJJ Bonvin, *Haddock versus haddock: new features and performance of haddock2. 0 on the capri targets*, Proteins: structure, function, and bioinformatics **69** (2007), no. 4, 726–733.

[38] Warren Lyford DeLano, *Pymol*, 2002.

[39] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, *Bert: Pre-training of deep bidirectional transformers for language understanding*, arXiv preprint arXiv:1810.04805 (2018).

[40] Todd J Dolinsky, Paul Czodrowski, Hui Li, Jens E Nielsen, Jan H Jensen, Gerhard Klebe, and Nathan A Baker, *Pdb2pqr: expanding and upgrading automated preparation of biomolecular structures for molecular simulations*, Nucleic acids research **35** (2007), no. suppl_2, W522–W525.

[41] James Dunbar, Konrad Krawczyk, Jinwoo Leem, Terry Baker, Angelika Fuchs, Guy Georges, Jiye Shi, and Charlotte M Deane, *Sabdab: the structural antibody database*, Nucleic acids research **42** (2013), no. D1, D1140–D1146.

[42] ———, *Sabdab: the structural antibody database*, Nucleic acids research **42** (2014), no. D1, D1140–D1146.

[43] Fatma-Elzahraa Eid, Mahmoud ElHefnawi, and Lenwood S Heath, *Denovo: virus-host sequence-based protein–protein interaction prediction*, Bioinformatics **32** (2016), no. 8, 1144–1150.

[44] Ahmed Elnaggar, Michael Heinzinger, Christian Dallago, Ghalia Rihawi, Yu Wang, Llion Jones, Tom Gibbs, Tamas Feher, Christoph Angerer, Martin Steinegger, et al., *Prottrans: towards cracking the language of life's code through self-supervised deep learning and high performance computing*, arXiv preprint arXiv:2007.06225 (2020).

[45] Richard Evans, Michael O'Neill, Alexander Pritzel, Natasha Antropova, Andrew W Senior, Timothy Green, Augustin Žídek, Russell Bates, Sam Blackwell, Jason Yim, et al., *Protein complex prediction with alphafold-multimer*, BioRxiv (2021).

[46] Michael J Feldhaus, Robert W Siegel, Lee K Opresko, James R Coleman, Jane M Weaver Feldhaus, Yik A Yeung, Jennifer R Cochran, Peter Heinzelman, David Colby, Jeffrey Swers, et al., *Flow-cytometric isolation of human antibodies from a nonimmune saccharomyces cerevisiae surface display library*, Nature biotechnology **21** (2003), no. 2, 163.

[47] S Jane Flint, Vincent R Racaniello, Glenn F Rall, Theodora Hatziioannou, and Anna Marie Skalka, *Principles of virology, volume 2: pathogenesis and control*, John Wiley & Sons, 2020.

[48] Ryan A Flynn, Kayvon Pedram, Stacy A Malaker, Pedro J Batista, Benjamin AH Smith, Alex G Johnson, Benson M George, Karim Majzoub, Peter W Villalta, Jan E Carette, et al., *Small rnas are modified with n-glycans and displayed on the surface of living cells*, Cell **184** (2021), no. 12, 3109–3124.

[49] Alex Fout, Jonathon Byrd, Basir Shariat, and Asa Ben-Hur, *Protein interface prediction using graph convolutional networks*, Advances in neural information processing systems, 2017, pp. 6530–6539.

[50] P Gainza, F Sverrisson, F Monti, E Rodolà, D Boscaini, MM Bronstein, and BE Correia, *Deciphering interaction fingerprints from protein molecular surfaces using geometric deep learning*, Nature Methods (2019), 1–9.

[51] Pablo Gainza, Freyr Sverrisson, Frederico Monti, Emanuele Rodola, D Boscaini, MM Bronstein, and BE Correia, *Deciphering interaction fingerprints from protein molecular surfaces using geometric deep learning*, Nature Methods **17** (2020), no. 2, 184–192.

[52] Octavian-Eugen Ganea, Xinyuan Huang, Charlotte Bunne, Yatao Bian, Regina Barzilay, Tommi Jaakkola, and Andreas Krause, *Independent se (3)-equivariant models for end-to-end rigid protein docking*, arXiv preprint arXiv:2111.07786 (2021).

[53] Chao Gao, Kathrin Stavenhagen, Barbara Eckmair, Tanya R McKitrick, Akul Y Mehta, Yasuyuki Matsumoto, Alyssa M McQuillan, Melinda S Hanes, Deniz Eris, Kelly J Baker, et al., *Differential recognition of oligomannose isomers by glycan-binding proteins involved in innate and adaptive immunity*, Science Advances **7** (2021), no. 24, eabf6834.

[54] Wilfredo F Garcia-Beltran, Kerri J St Denis, Angelique Hoelzemer, Evan C Lam, Adam D Nitido, Maegan L Sheehan, Cristhian Berrios, Onosereme Ofoman, Christina C Chang, Blake M Hauser, et al., *mrna-based covid-19 vaccine boosters induce neutralizing immunity against sars-cov-2 omicron variant*, Cell (2022).

[55] Christoph Geisler and Donald L Jarvis, *Letter to the glyco-forum: Effective glycoanalysis with maackia amurensis lectins requires a clear understanding of their binding specificities*, Glycobiology **21** (2011), no. 8, 988–993.

[56] Rafael Gómez-Bombarelli, Jennifer N Wei, David Duvenaud, José Miguel Hernández-Lobato, Benjamín Sánchez-Lengeling, Dennis Sheberla, Jorge Aguilera-Iparraguirre, Timothy D Hirzel, Ryan P Adams, and Alán Aspuru-Guzik, *Automatic chemical design using a data-driven continuous representation of molecules*, ACS central science **4** (2018), no. 2, 268–276.

[57] Karl E Griswold and Chris Bailey-Kellogg, *Design and engineering of deimmunized biotherapeutics*, Current opinion in structural biology **39** (2016), 79–88.

[58] Yuan Guo, Hadar Feinberg, Edward Conroy, Daniel A Mitchell, Richard Alvarez, Ola Blixt, Maureen E Taylor, William I Weis, and Kurt Drickamer, *Structural basis for distinct ligand-binding and targeting properties of the receptors dc-sign and dc-signr*, Nature structural & molecular biology **11** (2004), no. 7, 591–598.

[59] Erika Gustafsson, Anna Rosén, Karin Barchan, Kok PM van Kessel, Karin Haraldsson, Stina Lindman, Cecilia Forsberg, Lill Ljung, Karin Bryder, Björn Walse, et al., *Directed evolution of chemotaxis inhibitory protein of staphylococcus aureus generates biologically functional variants with reduced interaction with human antibodies*, Protein Engineering, Design & Selection **23** (2010), no. 2, 91–101.

[60] Brian B Haab and Zachary Klamer, *Advances in tools to determine the glycan-binding specificities of lectins and antibodies*, Molecular & Cellular Proteomics **19** (2020), no. 2, 224–232.

[61] Kosuke Hashimoto, Ichigaku Takigawa, Motoki Shiga, Minoru Kanehisa, and Hiroshi Mamitsuka, *Mining significant tree patterns in carbohydrate sugar chains*, Bioinformatics **24** (2008), no. 16, i167–i173.

[62] Pernille Haste Andersen, Morten Nielsen, and OLE Lund, *Prediction of residues in discontinuous b-cell epitopes using protein 3d structures*, Protein Science **15** (2006), no. 11, 2558–2567.

[63] Lu He, Alan M Friedman, and Chris Bailey-Kellogg, *A divide-and-conquer approach to determine the pareto frontier for optimization of protein engineering experiments*, Proteins: Structure, Function, and Bioinformatics **80** (2012), no. 3, 790–806.

[64] Steven Henikoff and Jorja G Henikoff, *Amino acid substitution matrices from protein blocks*, Proceedings of the National Academy of Sciences **89** (1992), no. 22, 10915–10919.

[65] Masae Hosoda, Yushi Takahashi, Masaaki Shiota, Daisuke Shinmachi, Renji Inomoto, Shinichi Higashimoto, and Kiyoko F Aoki-Kinoshita, *Mcaw-db: A glycan profile database capturing the ambiguity of glycan recognition patterns*, Carbohydrate research **464** (2018), 44–56.

[66] Casey K Hua, Albert T Gacerez, Charles L Sentman, Margaret E Ackerman, Yoonjoo Choi, and Chris Bailey-Kellogg, *Computationally-driven identification of antibody epitopes*, Elife **6** (2017), e29023.

[67] Ying Huang, Beifang Niu, Ying Gao, Limin Fu, and Weizhong Li, *Cd-hit suite: a web server for clustering and comparing biological sequences*, Bioinformatics **26** (2010), no. 5, 680–682.

[68] Jaime Huerta-Cepas, François Serra, and Peer Bork, *Ete 3: reconstruction, analysis, and visualization of phylogenomic data*, Molecular biology and evolution **33** (2016), no. 6, 1635–1638.

[69] Howook Hwang, Brian Pierce, Julian Mintseris, Joël Janin, and Zhiping Weng, *Protein–protein docking benchmark version 3.0*, Proteins: Structure, Function, and Bioinformatics **73** (2008), no. 3, 705–709.

[70] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al., *Spatial transformer networks*, Advances in neural information processing systems, 2015, pp. 2017–2025.

[71] Charles A Janeway, Paul Travers, Mark Walport, and Donald J Capra, *Immunobiology*, Taylor & Francis Group UK: Garland Science, 2001.

[72] Madeleine F Jennewein and Galit Alter, *The immunoregulatory roles of antibody glycosylation*, Trends in immunology **38** (2017), no. 5, 358–372.

[73] Martin Closter Jespersen, Swapnil Mahajan, Bjoern Peters, Morten Nielsen, and Paolo Marcatili, *Antibody specific b-cell epitope predictions: leveraging information from antibody-antigen protein complexes*, Frontiers in immunology **10** (2019), 298.

[74] Xiao-Chi Jia, Robert Raya, Li Zhang, Orit Foord, Wynn L Walker, Michael L Gallo, Mary Haak-Frendscho, Larry L Green, and C Geoffrey Davis, *A novel method of multiplexed competitive antibody binning for the characterization of monoclonal antibodies*, Journal of immunological methods **288** (2004), no. 1-2, 91–98.

[75] Lei Jin and James A Wells, *Dissecting the energetics of an antibody-antigen interface by alanine shaving and molecular grafting*, Protein Science **3** (1994), no. 12, 2351–2357.

[76] A Abragam Joseph, Alonso Pardo-Vargas, and Peter H Seeberger, *Total synthesis of polysaccharides by automated glycan assembly*, Journal of the American Chemical Society **142** (2020), no. 19, 8561–8564.

[77] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al., *Highly accurate protein structure prediction with alphafold*, Nature **596** (2021), no. 7873, 583–589.

[78] Florian Kiefer, Konstantin Arnold, Michael Künzli, Lorenza Bordoli, and Torsten Schwede, *The swiss-model repository and associated resources*, Nucleic acids research **37** (2009), no. suppl_1, D387–D392.

[79] Diederik P Kingma and Jimmy Ba, *Adam: A method for stochastic optimization*, arXiv preprint arXiv:1412.6980 (2014).

[80] Thomas N Kipf and Max Welling, *Semi-supervised classification with graph convolutional networks*, International Conference on Learning Representations (2017).

[81] Zachary Klamer and Brian Haab, *Automated identification of lectin fine specificities from glycan-array data*, Glycan-Based Cellular Communication: Techniques for Carbohydrate-Protein Interactions, ACS Publications, 2020, pp. 67–82.

[82] Zachary Klamer, Ben Staal, Anthony R Prudden, Lin Liu, David F Smith, Geert-Jan Boons, and Brian Haab, *Mining high-complexity motifs in glycans: A new language to uncover the fine specificities of lectins and glycosidases*, Analytical chemistry **89** (2017), no. 22, 12342–12350.

[83] Dima Kozakov, David R Hall, Bing Xia, Kathryn A Porter, Dzmitry Padhorny, Christine Yueh, Dmitri Beglov, and Sandor Vajda, *The cluspro web server for protein–protein docking*, Nature protocols **12** (2017), no. 2, 255–278.

[84] Konrad Krawczyk, Xiaofeng Liu, Terry Baker, Jiye Shi, and Charlotte M Deane, *Improving b-cell epitope prediction and its application to global antibody-antigen docking*, Bioinformatics **30** (2014), no. 16, 2288–2294.

[85] Jens Vindahl Kringelum, Claus Lundegaard, Ole Lund, and Morten Nielsen, *Reliable b cell epitope predictions: impacts of method development and improved benchmarking*, PLoS computational biology **8** (2012), no. 12, e1002829.

[86] Jack Kyte and Russell F Doolittle, *A simple method for displaying the hydropathic character of a protein*, Journal of molecular biology **157** (1982), no. 1, 105–132.

[87] Alan L Schmaljohn, *Protective antiviral antibodies that lack neutralizing activity: precedents and evolution of concepts*, Current HIV research **11** (2013), no. 5, 345–353.

[88] Michael C Lawrence and Peter M Colman, *Shape complementarity at protein/protein interfaces*, 1993.

[89] Marc F Lensink, Nurul Nadzirin, Sameer Velankar, and Shoshana J Wodak, *Modeling protein-protein, protein-peptide, and protein-oligosaccharide complexes: Capri 7th edition*, Proteins: Structure, Function, and Bioinformatics **88** (2020), no. 8, 916–938.

[90] Bo-Han Li and Xin-Shan Ye, *Recent advances in glycan synthesis*, Current Opinion in Chemical Biology **58** (2020), 20–27.

[91] Lei Li, Wanyi Guan, Gaolan Zhang, Zhigang Wu, Hai Yu, Xi Chen, and Peng G Wang, *Microarray analyses of closely related glycoforms reveal different accessibilities of glycan determinants on n-glycan branches*, Glycobiology **30** (2020), no. 5, 334–345.

[92] Yuqing Li, Dongqi Liu, Yating Wang, Wenquan Su, Gang Liu, and Weijie Dong, *The importance of glycans of viral and host proteins in enveloped virus infection*, Frontiers in Immunology **12** (2021), 1544.

[93] Zhouhan Lin, Minwei Feng, Cicero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio, *A structured self-attentive sentence embedding*, arXiv preprint arXiv:1703.03130 (2017).

[94] Wenhai Liu, Masanori Onda, Byungkook Lee, Robert J Kreitman, Raffit Hassan, Laiman Xiang, and Ira Pastan, *Recombinant immunotoxin engineered for low immunogenicity and antigenicity by identifying and silencing human b-cell epitopes*, Proceedings of the National Academy of Sciences **109** (2012), no. 29, 11782–11787.

[95] Romain Lopez, Adam Gayoso, and Nir Yosef, *Enhancing scientific discoveries in molecular biology with deep generative models*, Molecular Systems Biology **16** (2020), no. 9, e9198.

[96] Chafen Lu, Mazen Ferzly, Junichi Takagi, and Timothy A Springer, *Epitope mapping of antibodies to the c-terminal region of the integrin $\beta 2$ subunit reveals regions that become exposed upon receptor activation*, The Journal of Immunology **166** (2001), no. 9, 5629–5637.

[97] Lenette L Lu, Todd J Suscovich, Sarah M Fortune, and Galit Alter, *Beyond binding: antibody effector functions in infectious diseases*, Nature Reviews Immunology **18** (2018), no. 1, 46–61.

[98] Jarjapu Mahita, Dong-Gun Kim, Sumin Son, Yoonjoo Choi, Hak-Sung Kim, and Chris Bailey-Kellogg, *Computational epitope binning reveals functional equivalence of sequence-divergent paratopes*, Computational and structural biotechnology journal **20** (2022), 2169–2180.

[99] Daniel Maturana and Sebastian Scherer, *Voxnet: A 3d convolutional neural network for real-time object recognition*, 2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), IEEE, 2015, pp. 922–928.

[100] Leland McInnes, John Healy, and James Melville, *Umap: Uniform manifold approximation and projection for dimension reduction*, arXiv preprint arXiv:1802.03426 (2018).

[101] Joshua Meier, Roshan Rao, Robert Verkuil, Jason Liu, Tom Sercu, and Alex Rives, *Language models enable zero-shot prediction of the effects of mutations on protein function*, Advances in Neural Information Processing Systems **34** (2021), 29287–29303.

[102] D Mirelman, *Introduction to microbial lectins and agglutinins*, Microbial Lectins and Agglutins (1986), 1–19.

[103] Carter A Mitchell, Koreen Ramessar, and Barry R O'Keefe, *Antiviral lectins: Selective inhibitors of viral entry*, Antiviral research **142** (2017), 37–54.

[104] David C Montefiori, Lynn Morris, Guido Ferrari, and John R Mascola, *Neutralizing and other antiviral antibodies in hiv-1 infection and vaccination*, Current Opinion in HIV and AIDS **2** (2007), no. 3, 169.

[105] Gustavo Marçal Schmidt Garcia Moreira, Viola Fühner, and Michael Hust, *Epitope mapping by phage display*, Phage Display, Springer, 2018, pp. 497–518.

[106] Kelley W Moremen, Michael Tiemeyer, and Alison V Nairn, *Vertebrate protein glycosylation: diversity, synthesis and function*, Nature reviews Molecular cell biology **13** (2012), no. 7, 448–462.

[107] Yoichi Murakami and Kenji Mizuguchi, *Applying the naïve bayes classifier with kernel density estimation to the prediction of protein–protein interaction sites*, Bioinformatics **26** (2010), no. 15, 1841–1848.

[108] Alexey G Murzin, Steven E Brenner, Tim Hubbard, and Cyrus Chothia, *Scop: a structural classification of proteins database for the investigation of sequences and structures*, Journal of molecular biology **247** (1995), no. 4, 536–540.

[109] Satoshi Nagata and Ira Pastan, *Removal of b cell epitopes as a practical approach for reducing the immunogenicity of foreign protein-based therapeutics*, Advanced drug delivery reviews **61** (2009), no. 11, 977–985.

[110] Yoshiki Narimatsu, Hiren J Joshi, Rebecca Nason, Julie Van Coillie, Richard Karlsson, Lingbo Sun, Zilu Ye, Yen-Hsi Chen, Katrine T Schjoldager, Catharina Steentoft, et al., *An atlas of human glycosylation pathways enables display of the human glycome by gene engineered cells*, Molecular cell **75** (2019), no. 2, 394–407.

[111] Hani Neuvirth, Ran Raz, and Gideon Schreiber, *Promate: a structure based prediction program to identify the location of protein–protein binding sites*, Journal of molecular biology **338** (2004), no. 1, 181–199.

[112] Johan Nilvebrant and Johan Rockberg, *An introduction to epitope mapping*, Epitope Mapping Protocols, Springer, 2018, pp. 1–10.

[113] Wataru Nishima, Naoyuki Miyashita, Yoshiki Yamaguchi, Yuji Sugita, and Suyong Re, *Effect of bisecting glcnac and core fucosylation on conformational properties of biantennary complex-type n-glycans in solution*, The Journal of Physical Chemistry B **116** (2012), no. 29, 8504–8512.

[114] Utkan Ogmen, Ozlem Keskin, A Selim Aytuna, Ruth Nussinov, and Attila Gursoy, *Prism: protein interactions by structural matching*, Nucleic acids research **33** (2005), no. suppl_2, W331–W336.

[115] Tobias H Olsen, Fergus Boyles, and Charlotte M Deane, *Observed antibody space: A diverse database of cleaned, annotated, and translated unpaired and paired antibody sequences*, Protein Science **31** (2022), no. 1, 141–146.

[116] Masanori Onda, Richard Beers, Laiman Xiang, Byungkook Lee, John E Weldon, Robert J Kreitman, and Ira Pastan, *Recombinant immunotoxin against b-cell malignancies with no immunogenicity in mice by removal of b-cell epitopes*, Proceedings of the National Academy of Sciences **108** (2011), no. 14, 5742–5747.

[117] Masanori Onda, Richard Beers, Laiman Xiang, Satoshi Nagata, Qing-cheng Wang, and Ira Pastan, *An immunotoxin with greatly reduced immunogenicity by identification and removal of b cell epitopes*, Proceedings of the National Academy of Sciences **105** (2008), no. 32, 11311–11316.

[118] Masanori Onda, Satoshi Nagata, David J FitzGerald, Richard Beers, Robert J Fisher, James J Vincent, Byungkook Lee, Michihiro Nakamura, Jaulang Hwang, Robert J Kreitman, et al., *Characterization of the b cell epitopes associated with a truncated form of pseudomonas exotoxin (pe38) used to make immunotoxins for the treatment of cancer patients*, The Journal of Immunology **177** (2006), no. 12, 8822–8834.

[119] Robert Osada, Thomas Funkhouser, Bernard Chazelle, and David Dobkin, *Shape distributions*, ACM Transactions on Graphics (TOG) **21** (2002), no. 4, 807–832.

[120] Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli, *fairseq: A fast, extensible toolkit for sequence modeling*, arXiv preprint arXiv:1904.01038 (2019).

[121] Ralph Pantophlet and Dennis R Burton, *Gp120: target for neutralizing hiv-1 antibodies*, Annual review of immunology **24** (2006), no. 1, 739–769.

[122] Ira Pastan, *Immunotoxins containing pseudomonas exotoxin a: a short history*, Cancer Immunology, Immunotherapy **52** (2003), no. 5, 338.

[123] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al., *Pytorch: An imperative style, high-performance deep learning library*, arXiv preprint arXiv:1912.01703 (2019).

[124] Tony Pawson and Piers Nash, *Protein–protein interactions define specificity in signal transduction*, Genes & development **14** (2000), no. 9, 1027–1047.

[125] Brian G Pierce, Kevin Wiehe, Howook Hwang, Bong-Hyun Kim, Thom Vreven, and Zhiping Weng, *Zdock server: interactive docking prediction of protein–protein complexes and symmetric multimers*, Bioinformatics **30** (2014), no. 12, 1771–1773.

[126] Srivamshi Pittala and Chris Bailey-Kellogg, *Learning context-aware structural representations to predict antigen and antibody binding interfaces*, Bioinformatics (2020).

[127] Aleksey Porollo and Jarosław Meller, *Prediction-based fingerprints of protein–protein interactions*, Proteins: Structure, Function, and Bioinformatics **66** (2007), no. 3, 630–645.

[128] Andrew Porter, Tingting Yue, Lee Heeringa, Steven Day, Edward Suh, and Brian B Haab, *A motif-based analysis of glycan array data to determine the specificities of glycan-binding proteins*, Glycobiology **20** (2010), no. 3, 369–380.

[129] Robert Porzig, David Singer, and Ralf Hoffmann, *Epitope mapping of mabs at8 and tau5 directed against hyperphosphorylated regions of the human tau protein*, Biochemical and biophysical research communications **358** (2007), no. 2, 644–649.

[130] Kathleen P Pratt, *Anti-drug antibodies: emerging approaches to predict, reduce or reverse biotherapeutic immunogenicity*, Antibodies **7** (2018), no. 2, 19.

[131] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas, *Pointnet: Deep learning on point sets for 3d classification and segmentation*, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 652–660.

[132] Kenneth E Quickel Jr, Richard Selden, Jacques R Caldwell, Nedo F Nora, and William Schaffner, *Efficacy and safety of topical lysostaphin treatment of persistent nasal carriage of staphylococcus aureus*, Applied microbiology **22** (1971), no. 3, 446–450.

[133] Jennifer L Remmel, Kathryn S Beauchemin, Akaash K Mishra, Julia C Frei, Jonathan R Lai, Chris Bailey-Kellogg, and Margaret E Ackerman, *Combinatorial resurfacing of dengue envelope protein domain iii antigens selectively ab-*

lates epitopes associated with serotype-specific or infection-enhancing antibody responses, ACS Combinatorial Science **22** (2020), no. 9, 446–456.

[134] Nicolas Renaud, Cunliang Geng, Sonja Georgievska, Francesco Ambrosetti, Lars Ridder, Dario F Marzella, Manon F Réau, Alexandre MJJ Bonvin, and Li C Xue, *Deeprank: a deep learning framework for data mining 3d protein-protein interfaces*, Nature communications **12** (2021), no. 1, 1–8.

[135] Alexander Rives, Joshua Meier, Tom Sercu, Siddharth Goyal, Zeming Lin, Jason Liu, Demi Guo, Myle Ott, C Lawrence Zitnick, Jerry Ma, et al., *Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences*, Proceedings of the National Academy of Sciences **118** (2021), no. 15.

[136] Nimrod D Rubinstein, Itay Mayrose, Dan Halperin, Daniel Yekutieli, Jonathan M Gershoni, and Tal Pupko, *Computational characterization of b-cell epitopes*, Molecular immunology **45** (2008), no. 12, 3477–3489.

[137] Ruben Sanchez-Garcia, Carlos Oscar Sánchez Sorzano, José María Carazo, and Joan Segura, *Bipspi: a method for the prediction of partner-specific protein–protein interfaces*, Bioinformatics **35** (2019), no. 3, 470–477.

[138] Huub Schellekens, *Immunogenicity of therapeutic proteins: clinical implications and future prospects*, Clinical therapeutics **24** (2002), no. 11, 1720–1740.

[139] Joerg U Schmohl, Deborah Todhunter, Seung Oh, and Daniel A Vallera, *Mutagenic deimmunization of diphtheria toxin for use in biologic drug development*, Toxins **7** (2015), no. 10, 4067–4082.

[140] Michael Schneider, Esam Al-Shareffi, and Robert S Haltiwanger, *Biological functions of fucose in mammals*, Glycobiology **27** (2017), no. 7, 601–618.

[141] Dina Schneidman-Duhovny, Yuval Inbar, Ruth Nussinov, and Haim J Wolfson, *Patchdock and symmdock: servers for rigid and symmetric docking*, Nucleic acids research **33** (2005), no. suppl_2, W363–W367.

[142] Inbal Sela-Culang, Mohammed Rafii-El-Idrissi Benhnia, Michael H Matho, Thomas Kaever, Matt Maybeno, Andrew Schlossman, Guy Nimrod, Sheng Li, Yan Xiang, Dirk Zajonc, et al., *Using a combined computational-experimental approach to predict antibody-specific b cell epitopes*, Structure **22** (2014), no. 4, 646–657.

[143] Inbal Sela-Culang, Vered Kunik, and Yanay Ofran, *The structural basis of antibody-antigen recognition*, Frontiers in immunology **4** (2013), 302.

[144] Inbal Sela-Culang, Yanay Ofran, and Bjoern Peters, *Antibody specific epitope prediction—emergence of a new paradigm*, Current opinion in virology **11** (2015), 98–102.

[145] Swaminathan Sethu, Karthik Govindappa, Mohammad Alhaidari, Munir Pirmohamed, Kevin Park, and Jean Sathish, *Immunogenicity to biologics: mechanisms, prediction and reduction*, Archivum immunologiae et therapiae experimentalis **60** (2012), no. 5, 331–344.

[146] Benjamin A Shoemaker and Anna R Panchenko, *Deciphering protein–protein interactions. part i. experimental techniques and databases*, PLoS computational biology **3** (2007), no. 3, e42.

[147] Kresimir Sikic, Sanja Tomic, and Oliviero Carugo, *Systematic comparison of crystal and nmr protein structures deposited in the protein data bank*, The open biochemistry journal **4** (2010), no. 1.

[148] Aroop Sircar and Jeffrey J Gray, *Snugdock: paratope structural optimization during antibody-antigen docking compensates for errors in antibody homology models*, PloS computational biology **6** (2010), no. 1, e1000644.

[149] Arvind Sivasubramanian, Patricia Estep, Heather Lynaugh, Yao Yu, Adam Miles, Josh Eckman, Kevin Schutz, Crystal Piffath, Nadthakarn Boland, Rebecca Hurley Niles, et al., *Broad epitope coverage of a human in vitro antibody library*, MAbs, vol. 9, Taylor & Francis, 2017, pp. 29–42.

[150] Vadim Slynko, Mario Schubert, Shin Numao, Michael Kowarik, Markus Aebi, and Frédéric H-T Allain, *Nmr structure determination of a segmentally labeled glycoprotein using in vitro glycosylation*, Journal of the American Chemical Society **131** (2009), no. 3, 1274–1281.

[151] Xuezheng Song, Yi Lasanajak, Baoyun Xia, Jamie Heimburg-Molinaro, Jeanne M Rhea, Hong Ju, Chunmei Zhao, Ross J Molinaro, Richard D Cummings, and David F Smith, *Shotgun glycomics: a microarray strategy for functional glycomics*, Nature methods **8** (2011), no. 1, 85–90.

[152] Tanlin Sun, Bo Zhou, Luhua Lai, and Jianfeng Pei, *Sequence-based prediction of protein protein interaction using a deep-learning algorithm*, BMC bioinformatics **18** (2017), no. 1, 1–8.

[153] Joel L Sussman, Dawei Lin, Jiansheng Jiang, Nancy O Manning, Jaime Prilusky, Otto Ritter, and Enrique E Abola, *Protein data bank (pdb): database of three-dimensional structural information of biological macromolecules*, Acta Crystallographica Section D: Biological Crystallography **54** (1998), no. 6, 1078–1084.

[154] Freyr Sverrisson, Jean Feydy, Joshua Southern, Michael M Bronstein, and Bruno Correia, *Physics-informed deep neural network for rigid-body protein docking*, ICLR2022 Machine Learning for Drug Discovery, 2022.

[155] Wilson L Taylor, *"cloze procedure": A new tool for measuring readability*, Journalism quarterly **30** (1953), no. 4, 415–433.

[156] Raphael Townshend, Rishi Bedi, Patricia Suriana, and Ron Dror, *End-to-end learning on 3d protein structure for interface prediction*, Advances in Neural Information Processing Systems, 2019, pp. 15642–15651.

[157] Rakesh Trivedi and Hampapathalu Adimurthy Nagarajaram, *Amino acid substitution scoring matrices specific to intrinsically disordered regions in proteins*, Scientific reports **9** (2019), no. 1, 1–12.

[158] Hannah L Turner, Raiees Andrabi, Christopher A Cottrell, Sara T Richey, Ge Song, Sean Callaghan, Fabio Anzanello, Tyson J Moyer, Wuhbet Abraham, Mariane Melo, et al., *Disassembly of hiv envelope glycoprotein trimer immunogens is driven by antibodies elicited via immunization*, bioRxiv (2021).

[159] Sandor Vajda, Christine Yueh, Dmitri Beglov, Tanggis Bohnuud, Scott E Mottarella, Bing Xia, David R Hall, and Dima Kozakov, *New additions to the c lus p ro server motivated by capri*, Proteins: Structure, Function, and Bioinformatics **85** (2017), no. 3, 435–444.

[160] Ajit Varki, Richard D Cummings, Markus Aebi, Nicole H Packer, Peter H Seeberger, Jeffrey D Esko, Pamela Stanley, Gerald Hart, Alan Darvill, Taroh Kinoshita, et al., *Symbol nomenclature for graphical representations of glycans*, Glycobiology **25** (2015), no. 12, 1323–1324.

[161] Ajit Varki and Pascal Gagneux, *Biological functions of glycans*, Essentials of Glycobiology [Internet]. 3rd edition. (Ajit Varki, Richard D Cummings, Jeffrey D Esko, Pamela Stanley, Gerald W Hart, Markus Aebi, Alan G Darvill, Taroh Kinoshita, Nicolle H Packer, James H Prestegard, et al., eds.), Cold Spring Harbor Laboratory Press, Cold Spring Harbor (NY), 2017.

[162] Ajit Varki and Stuart Kornfeld, *Historical background and overview*, Essentials of Glycobiology [Internet]. 3rd edition. (Ajit Varki, Richard D Cummings, Jeffrey D Esko, Pamela Stanley, Gerald W Hart, Markus Aebi, Alan G Darvill, Taroh Kinoshita, Nicolle H Packer, James H Prestegard, et al., eds.), Cold Spring Harbor Laboratory Press, Cold Spring Harbor (NY), 2017.

[163] Ajit Varki, Ronald L Schnaar, and Roland Schauer, *Sialic acids and other nonulosonic acids*, Essentials of Glycobiology [Internet]. 3rd edition. (Ajit Varki, Richard D Cummings, Jeffrey D Esko, Pamela Stanley, Gerald W Hart, Markus Aebi, Alan G Darvill, Taroh Kinoshita, Nicolle H Packer, James H Prestegard, et al., eds.), Cold Spring Harbor Laboratory Press, Cold Spring Harbor (NY), 2017.

[164] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin, *Attention is all you need*, arXiv preprint arXiv:1706.03762 (2017).

[165] Deeptak Verma, Gevorg Grigoryan, and Chris Bailey-Kellogg, *Structure-based design of combinatorial mutagenesis libraries*, Protein science **24** (2015), no. 5, 895–908.

[166] Oriol Vinyals, Samy Bengio, and Manjunath Kudlur, *Order matters: Sequence to sequence for sets*, arXiv preprint arXiv:1511.06391 (2015).

[167] Lane Votapka and Rommie E Amaro, *Multistructural hot spot characterization with ftprod*, Bioinformatics **29** (2013), no. 3, 393–394.

[168] Thom Vreven, Iain H Moal, Anna Vangone, Brian G Pierce, Panagiotis L Kastritis, Mieczyslaw Torchala, Raphael Chaleil, Brian Jiménez-García, Paul A Bates, Juan Fernandez-Recio, et al., *Updates to the integrated protein–protein interaction benchmarks: docking benchmark version 5 and affinity benchmark version 2*, Journal of molecular biology **427** (2015), no. 19, 3031–3041.

[169] Renxiao Wang, Xueliang Fang, Yipin Lu, Chao-Yie Yang, and Shaomeng Wang, *The pdbbind database: methodologies and updates*, Journal of medicinal chemistry **48** (2005), no. 12, 4111–4119.

[170] Xiao Wang, Genki Terashi, Charles W Christoffer, Mengmeng Zhu, and Daisuke Kihara, *Protein docking model evaluation by 3d deep convolutional neural networks*, Bioinformatics **36** (2020), no. 7, 2113–2118.

[171] Zhen Wang, Zoeisha S Chinoy, Shailesh G Ambre, Wenjie Peng, Ryan McBride, Robert P de Vries, John Glushka, James C Paulson, and Geert-Jan Boons, *A general strategy for the chemoenzymatic synthesis of asymmetrically branched n-glycans*, Science **341** (2013), no. 6144, 379–383.

[172] Brian D Weitzner, Jeliazko R Jeliazkov, Sergey Lyskov, Nicholas Marze, Daisuke Kuroda, Rahel Frick, Jared Adolf-Bryfogle, Naireeta Biswas, Roland L Dunbrack Jr, and Jeffrey J Gray, *Modeling and docking of antibody structures with rosetta*, Nature protocols **12** (2017), no. 2, 401.

[173] Xueling Wu, Zhi-Yong Yang, Yuxing Li, Carl-Magnus Hogerkorp, William R Schief, Michael S Seaman, Tongqing Zhou, Stephen D Schmidt, Lan Wu, Ling

Xu, et al., *Rational design of envelope identifies broadly neutralizing human monoclonal antibodies to hiv-1*, Science **329** (2010), no. 5993, 856–861.

[174] Chunming Xu and Scott A Jackson, *Machine learning and complex biological data*, 2019.

[175] JY Xu, MK Gorny, T Palker, S Karwowska, and S Zolla-Pazner, *Epitope mapping of two immunodominant domains of gp41, the transmembrane protein of human immunodeficiency virus type 1, using ten human monoclonal antibodies*, Journal of virology **65** (1991), no. 9, 4832–4838.

[176] Yumeng Yan, Huanyu Tao, Jiahua He, and Sheng-You Huang, *The hdock server for integrated protein–protein docking*, Nature protocols **15** (2020), no. 5, 1829–1852.

[177] Yumeng Yan, Di Zhang, Pei Zhou, Botong Li, and Sheng-You Huang, *Hdock: a web server for protein–protein and protein–dna/rna docking based on a hybrid strategy*, Nucleic acids research **45** (2017), no. W1, W365–W373.

[178] Jingyun Yang, Wei Wang, Zimin Chen, Shuaiyao Lu, Fanli Yang, Zhenfei Bi, Linlin Bao, Fei Mo, Xue Li, Yong Huang, et al., *A vaccine targeting the rbd of the s protein of sars-cov-2 induces protective immunity*, Nature **586** (2020), no. 7830, 572–577.

[179] Sharon Yehuda and Vered Padler-Karavani, *Glycosylated biotherapeutics: immunological effects of n-glycolylneuraminic acid*, Frontiers in immunology **11** (2020), 21.

[180] Hong Zeng, Sheng Wang, Tianming Zhou, Feifeng Zhao, Xiufeng Li, Qing Wu, and Jinbo Xu, *Complexcontact: a web server for inter-protein contact prediction using deep learning*, Nucleic acids research **46** (2018), no. W1, W432–W437.

[181] Min Zeng, Fuhao Zhang, Fang-Xiang Wu, Yaohang Li, Jianxin Wang, and Min Li, *Protein–protein interaction site prediction through combining local and global features with deep neural networks*, Bioinformatics **36** (2020), no. 4, 1114–1120.

[182] Fred Zepp, *Principles of vaccine design—lessons from nature*, Vaccine **28** (2010), C14–C24.

[183] Qiangfeng Cliff Zhang, Donald Petrey, Lei Deng, Li Qiang, Yu Shi, Chan Aye Thu, Brygida Bisikirska, Celine Lefebvre, Domenico Accili, Tony Hunter, et al., *Structure-based prediction of protein–protein interactions on a genome-wide scale*, Nature **490** (2012), no. 7421, 556.

[184] Hongliang Zhao, Seth A Brooks, Susan Eszterhas, Spencer Heim, Liang Li, Yan Q Xiong, Yongliang Fang, Jack R Kirsch, Deeptak Verma, Chris Bailey-Kellogg, et al., *Globally deimmunized lysostaphin evades human immune surveillance and enables highly efficacious repeat dosing*, Science advances **6** (2020), no. 36, eabb9011.

[185] Hongliang Zhao, Deeptak Verma, Wen Li, Yoonjoo Choi, Christian Ndong, Steven N Fiering, Chris Bailey-Kellogg, and Karl E Griswold, *Depletion of t cell epitopes in lysostaphin mitigates anti-drug antibody response and enhances antibacterial efficacy in vivo*, Chemistry & biology **22** (2015), no. 5, 629–639.

[186] Chen Zhou, Zikun Chen, Lu Zhang, Deyu Yan, Tiantian Mao, Kailin Tang, Tianyi Qiu, and Zhiwei Cao, *Seppa 3.0—enhanced spatial epitope prediction enabling glycoprotein antigens*, Nucleic acids research **47** (2019), no. W1, W388–W394.

[187] Léa V Zinsli, Noël Stierlin, Martin J Loessner, and Mathias Schmelcher, *Deimmunization of protein therapeutics–recent advances in experimental and compu-*

tational epitope prediction and deletion, Computational and Structural Biotechnology Journal **19** (2021), 315–329.