DIPLOMA THESIS

Antibody structure prediction using a sequence-only data set

carried out at

Institut for Protein Design
University of Washington, Seattle

under the supervision of

University of Washington:
David Baker

and

Technische Universität Wien:
Martina Marchetti-Deschmann

by

Benedikt Singer

Vienna on November 20, 2023

Kurzfassung

Die Vorhersage der Struktur von Antikörper-Antigen-Komplexen bleibt eine schwierige Aufgabe, da die komplementäritätsbestimmenden Regionen (engl. complementarity determining regions - CDR), die die Interaktion des Antikörpers bestimmen, einzigartig und unstrukturiert sind. Außerdem sind wenig koevolutionären Informationen über die Antikörper vorhanden, was die Strukturvorhersage zu einem schwierigen Problem macht. Eine zusätzliche Herausforderung in dieser Hinsicht stellt der Mangel an Strukturen von Antikörpern in der Proteindatenbank (PDB) dar. Die Fülle an experimentellen Daten zu Antikörpern, die nicht die Struktur betreffen, bietet jedoch eine vielversprechende Ressource für die Erforschung, indem diese Daten in die Strukturvorhersage integriert werden können. Eine Art von Daten sind experimentelle Bindungsdaten, die eine zusätzliche Dimension zur Integration und schließlich zur Unterstützung der Strukturvorhersage bieten.
Die Hypothese für dieses Projekt bestand darin, diese Art von Daten in ein neuronales Netzwerk zur Strukturvorhersage einzubeziehen und zu untersuchen, ob die korrekte Vorhersage einer spezifischen Antigen-Antikörper-Interaktion auch die Gesamtstrukturvorhersage verbessern kann; ein Problem von enormem pharmakologischen Interesse. Konkret haben wir eine Bibliothek von Antikörpern verwendet, von denen bekannt ist, dass sie das Spike-Protein von SARS-CoV-2 binden, einschließlich Informationen darüber, ob der Antikörper an die mutierte Rezeptorbindungsdomäne (RBD) des Proteins bindet. Dieser Ansatz der Zusammenführung von Sequenzdaten und Bindungsdaten wurde unseres Wissens bisher noch nicht angewandt. Wir haben uns überlegt, dass die Erkenntnis, ob ein Antikörper an eine bestimmte RBD-Mutante bindet oder nicht, dem Modell ein besseres strukturelles Verständnis von Antigen-Antikörper-Komplexen abverlangt und damit die Strukturvorhersage innerhalb eines Netzwerks verbessern könnte. Ziel dieser Arbeit ist es, die mit den Daten verbundenen Herausforderungen und Verzerrungen zu beschreiben und zu ermitteln, wie man mit diesen Daten trainieren kann.
Das Endergebnis des Projekts war, dass sich die Ableitung dieser Informationen über die Bindungsschnittstelle als kritisch erwies, die aus Daten des Deep Mutational Scanning (DMS) abgeleitet wurden. Die Ergebnisse deuten darauf hin, dass die Vorhersage von Antikörperstrukturen und Antikörper-Antigen-Bindungsschnittstellen wenig bis gar nicht verbessert werden konnte und eine anspruchsvolle Aufgabe bleibt. Zukünftige Ansätze können eine bessere Verfeinerung der Datensätze und eine verbesserte Ableitung der Informationen über die Bindungsschnittstelle umfassen. Darüber hinaus ist es wichtig zu betonen, dass die Diversifizierung von Trainingsdatensätzen und die Einbeziehung von Antikörpersequenzdaten von mehreren Zielproteinen einige der in dieser Arbeit entdeckten Probleme entschärfen kann.

Abstract

Antibody-antigen complex structure prediction remains a challenging task due to the unique unstructured nature inherent to the complementarity-determining regions (CDR) which determine the interaction of the antibody. Furthermore, co-evolutionary information on antibodies is sparse which makes it a challenging problem for structure prediction. An additional challenge on that aspect is posed by the lack of solved structures containing antibodies in the Protein Database (PDB). However, the abundance of non-structural experimental data on antibodies offers a promising resource for exploration by potentially integrating this data into structure prediction. One kind of data is experimental binding data which offers an additional dimension to integrate and eventually aid the structure prediction.
The hypothesis for this project was to incorporate this type of data into a structure-prediction neural network and investigate whether correctly predicting a specific target-antibody interaction can also improve overall structure prediction; a problem of enormous pharmacological interest. Specifically, we incorporated a library of antibodies known to bind the spike protein of SARS-CoV-2, including information on whether the antibody is binding to the mutated receptor binding domain (RBD) of the protein. This approach of amalgamation of sequence data and binding data has not, to our knowledge, been done previously. We reasoned that to learn whether or not an antibody would bind to a particular RBD mutant would require the model to learn a stronger structural understanding of antigen-antibody complexes, and thereby could improve structure prediction within a network. This thesis aims to describe and identify the challenges and biases associated with the data and how to train on it.
The final outcome of the project was that generating privileged information about the binding interface proved to be critical, which was derived from deep mutational scanning (DMS) data. The findings suggest that little to no improvement in the prediction of antibody structures and antibody antigen binding interfaces could be achieved and it remains a challenging task. Future work includes better data set refinement and improved derivation for the privileged information. Moreover, it is important to emphasize that diversifying training data sets and incorporating antibody sequence data from multiple targets can mitigate some of the issues discovered in this thesis.

Acknowledgements

First and foremost, I extend my deepest gratitude to Joe Watson and Nate Bennett for their invaluable guidance and mentorship during my time at the IPD in Seattle. Their extensive knowledge in protein structure prediction and their introduction to the world of protein design and science in general has been truly enlightening and them taking the time to share it with me is extremely appreciated. I am also deeply indebted to David Baker for generously allowing me to work in the lab and providing me with the opportunity to contribute to this significant project thus shaping my academic future in a significant way. I am grateful to Martina for her continual supervision in Austria and her unwavering support with all administrative matters.
A special acknowledgment goes to the wonderful people and amazing scientists at the IPD, whose warmth and openness created an amazing and unique learning environment. Their readiness to answer all my questions has been immensely helpful throughout my stay there. Of these people I would like to emphasize the following individuals in particular: David Juergens, DéJenaé See, Doug Tischer, Edin Muratspahic, Ellen Shrock, Florian Praetorius, Florence Hardy, Ivan Anishchenko, Jacob Gershon, Jue Wang, Linna An, Lisa Brandenburg, Ljubica Mihaljevic, Long Tran, Luki Goldschmidt, Lucas Arnoldt, Magnus Bauer, Paul Kim, Pascal Sturmfels, Ryan McHugh, Thomas Schlichthärle, and Zari Magness.

I also am grateful to my friends in Seattle for making my stay in the city a memorable and enjoyable experience, providing me with moments of joy and respite. I am especially thankful to Lisa for her unwavering friendship and unwavering support during my time in Seattle, helping me navigate life's challenges. Furthermore, I would like to extend my gratitude to Anna, whose encouragement led me to go on this journey to Seattle - and for last-minute proofreading the thesis. I also want to thank Alex, for thoroughly proof-reading this manuscript for me.
Ich schulde meiner gesamten Familie Dank für ihre unerschütterliche Ermutigung während meines Studiums. Ihre fortwährende Unterstützung meiner Lebensentscheidungen ist eine sehr wichtige Quelle von Stärke für mich.
I would also like to express my appreciation to the Austrian Marshall Plan and the TU Wien for their generous support through the funding. Thus, I also want to thank the Austrian taxpayers for providing the necessary financial backing for these institutions.

Eidesstattliche Erklärung

Ich erkläre an Eides statt, dass ich die vorliegende Diplomarbeit selbstständig und ohne fremde Hilfe verfasst, andere als die angegebenen Quellen und Hilfsmittel nicht benutzt bzw. die wörtlich oder sinngemäß entnommenen Stellen als solche kenntlich gemacht habe.

Wien, am November 20, 2023

Benedikt Singer

Contents

1 Introduction

The following chapters are intended to give an overview of the field of protein structure prediction in a historic context and the impact of the advent of deep learning. Especially the aspect of incorporating sequence-only data will be highlighted. Additionally, this chapter aims to give a small introduction into protein design and further, antibody design, as structure prediction is of clear relevance and importance to protein design. Furthermore, the text will delve into the clinical significance of antibodies and their potential as targeting drugs.

1.1 Protein Structure Prediction

In 1972 Christian Anfinsen postulated Anfinsen's Dogma, for which he was awarded the Nobel prize. This dogma lays the conceptual foundation for understanding protein structure and folding principles. This fundamental principle postulates that the native three-dimensional structure of a protein is inherently determined by its amino acid sequence. In other words, a protein's primary structure, represented by its sequence of amino acids, governs its unique and intricate three-dimensional fold which according to the hypothesis also represents the lowest energy state. Anfinsen's Dogma underscores the inherent relationship between a protein's sequence and its final folded structure on an energy landscape, establishing a guiding principle crucial to the field of protein structure prediction [1].

One of the first programs used for computational protein structure prediction was Rosetta which has been developed over the past two decades. The development of Rosetta began in the late 1990s, when David Baker and his colleagues at the University of Washington began working on a computational method for predicting protein structures. The initial version of Rosetta was released in the 2000s, and since then, the software has undergone many updates and improvements [2]. The protein structure prediction part of the software suite can be assigned to the physics-based structure prediction programs since the optimization of the structure is based on the minimization of an energy function which consists of various terms which aim to reflect the physics of protein folding [3].

When AlphaFold (AF1) conceived by DeepMind, made its debut and was showcased at the 13th Critical Assessment of Structure Prediction (CASP13) conference [4] [5], it exhibited noteworthy advancements compared to other structure prediction methods [6]. AF uses a deep learning network to predict protein structures, and has introduced numerous novelties like processing inter-residue geometries, and feeding the resultant coordinates alongside the evolutionary information conserved in homologous sequences into the network. After that, a subsequent energy minimization was conducted using gradient descent. In the case of the later published trRosetta, an acronym for transform-restrained Rosetta [7] - Rosetta was employed. Evolutionary information in the form of multiple sequence alignments (MSAs) have been used in models which were template-based since the first CASP conference in 1994, based on the

assumption that sequence based information is evolutionarily conserved and reflects into the structure [6] [8].

The most significant breakthrough occurred with the unveiling of a reengineered version of AlphaFold, termed AlphaFold2 (AF2) [9], at the 14th CASP conference in 2020. AlphaFold2 not only retained the established utilization of evolutionary information and data from templates — fragments derived from analogous proteins — but also incorporated these elements into a novel, transformer-based architecture. Among its innovations, AlphaFold2 stands out as an SE(3)-equivariant neural network, a feature which generally speaking increases performance and decreases training time.
To give a very broad overview, SE(3)-equivariance means that the 3D representation of the coordinates for the model remains consistent even when subjected to spatial transformations involving rotations and translations. This property decreases a network's necessary parameters, making it more efficient during training and runtime [10]. The success of this methodology prompted the subsequent development of RoseTTAfold, which adopted a similar strategy. Notably, RoseTTAfold introduced three distinct tracks for representing the evolutionary information in the first and the other two for the coordinates derived from evolutionary information called templates, as detailed in section 2.1. Furthermore, it also exploited the principles of SE(3)-equivariance.

In the subsequent years, there has been a transformative shift in the field of structural biology. DeepMind's continuous refinement of AlphaFold2 and the collaborative efforts of various research groups have led to enhancements in the network's performance tailored for specific tasks [11][12][13][14]. This evolution has positioned AlphaFold2 as an indispensable tool for experimental structural biologists seeking to unravel more complex structures. Notably, the AlphaFold Database, housing over 214 million predicted protein structures as of 16.11.23, has become a valuable resource, fostering research across diverse domains [15].
Subsequent releases this year, namely RoseTTAfold-NucleicAcid (RFNA) and RoseTTAfold All-Atom (RFAA) [16], have expanded the scope of structure prediction to encompass biological assemblies, incorporating proteins, nucleic acids, small molecules, metals, and covalent modifications. Furthermore, DeepMind and Isomorphic Labs released some vague results about the latest version of AlphaFold, called AlphaFold-latest [17]. This model, while capable of the same all-atom awareness as the methods mentioned earlier, distinguishes itself with significantly improved accuracy in antibody structure prediction. Notably, addressing the challenges in this area is a focus of ongoing work (manuscript in preparation) and improved by the approach pursued in this thesis.

1.2 Antibody Structure Prediction using fine-tuned RoseTTAfold2

As already mentioned general protein structure prediction [18] [9] shows high accuracy and reliability in predicting the structure of a single domain protein based on its respective amino acid sequence. Moreover, AlphaFold-Multimer was also published and extends the capabilities of AlphaFold2 to predict the structure of a protein complex, and together with RF2 it is considered the de facto gold standard in protein structure prediction.

IgFold [19], EquiFold [20] and ImmuneBuilder [21] are models which are specifically designed to predict the structure of antibodies. These models have shown improvement compared to AF2's predictions by giving these models more details about the specificity of antibodies in a fine-tuning step. However, the improvements observed with these models are relatively modest when compared to AF-Multimer. For instance, ABodyBuilder2, one of the models within the ImmuneBuilder suite, demonstrates only a $0.09\,\text{Å}$ enhancement in the root mean square deviation (RMSD) of a binding-specific region of an antibody compared to AF-Multimer. This overall indicates that the mentioned models are not yet suitable for the task at hand thus making it an unsolved challenge. Given this marginal difference in performance and considering the challenges encountered, the task of standalone antibody structure prediction appears to be constrained by limited data availability.
As discussed in section 1.3, antibodies present a unique challenge due to their highly loop-rich nature and the lack of a specific secondary structure associated with the CDRs. Moreover, there exists a lack of homologous sequences for antibodies thereby limiting the amount of evolutionary information available for training. Consequently, this makes it very hard for the current state-of-the-art structure prediction methods to correctly predict the structure of an antibody. The authors of (manuscript in preparation) set out to solve this task by coming up with a model specialized for predicting antibodies or antibody-antigen complexes which they call RFantibody. The origin of this model is the fine-tuned RF2 network on a data set which specifically consists of antibody structures with annotated CDRs originating from a data set called The Structural Antibody Database (SAbDab) [22][23]. Furthermore, RFantibody was fine-tuned using a curated data set of peptide-major histocompatibility complex (MHC) complexes bound to T-cell receptors (TCRs) which was manually curated [24].

1.3 Clinical Relevance of Antibodies

Broadly speaking antibodies represent proteins generated by the immune system capable of recognizing and binding to specific targets, such as viruses, bacteria, and cancer cells. Antibodies have become an important class of drugs in recent years due to their high specificity and low adverse effects [25]. As of June 2022 there were 163 approved antibody drugs for the treatment of different human diseases, encompassing cancer, autoimmune disorders, and chronic inflammatory diseases [26]. Additionally, to date, the global market size for monoclonal antibodies (mAbs) alone was estimated at approximately $\$210$ bn. in 2022 [27][25].
Figure 1.1 depicts the different ways antibodies or antibodies as drugs can interact with human cells or the human immune system. They are an attractive target as therapeutics for a wide range of reasons:

- High specificity and affinity: Antibodies have been selected for their ability to bind with high specificity and affinity to a wide variety of molecules. This allows them to target specific disease-causing agents, while minimizing off-target effects [28].
- Low toxicity: In therapy, antibodies are favored due to their low toxicity, which makes them beneficial for the treatment of cancer and autoimmune diseases. Their mechanism of action, which often involves blocking or modulating specific interactions, can be tuned to achieve the desired therapeutic effect with minimal side effects [28].

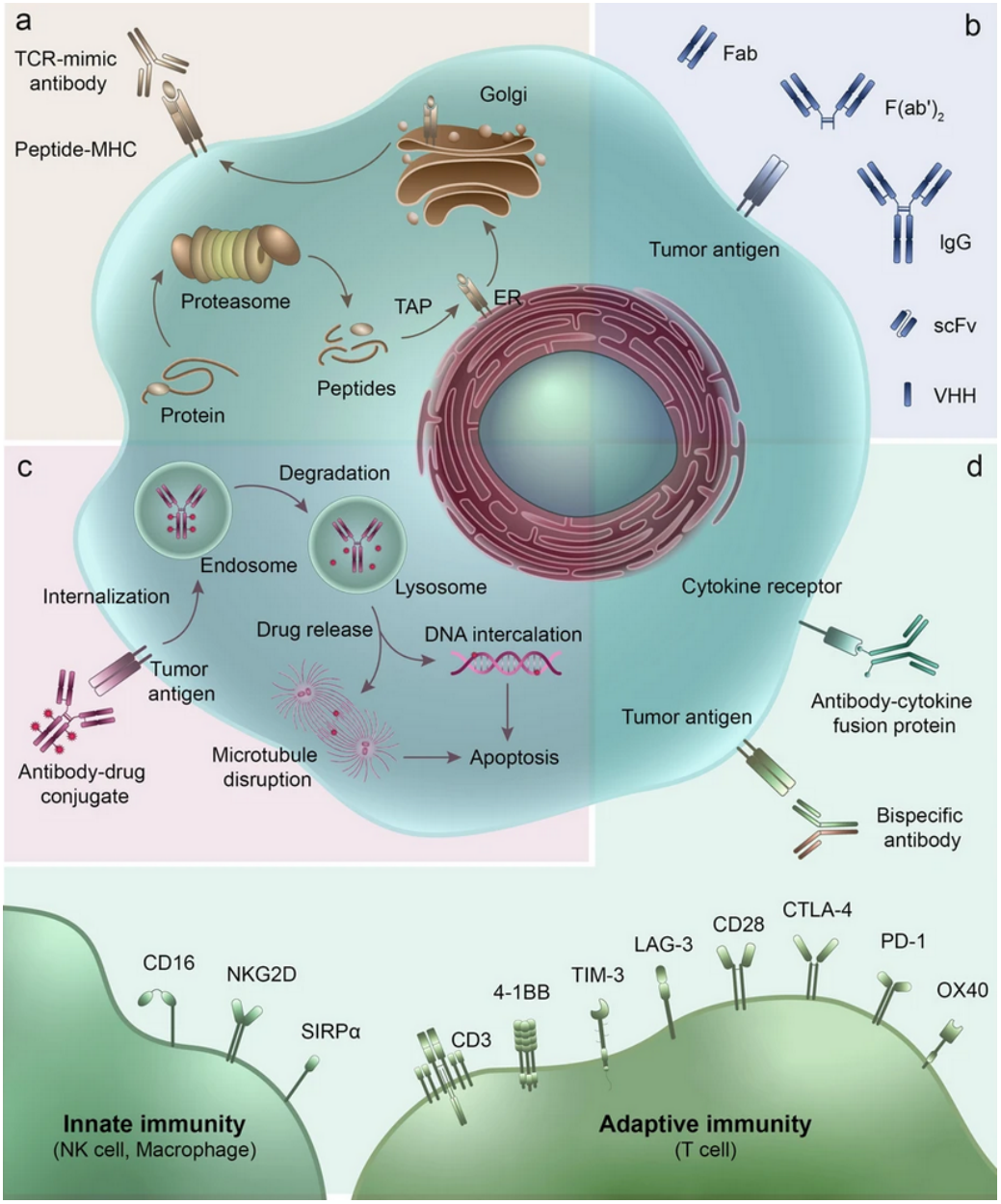


Figure 1.1: This figure by [28] gives a comprehensive overview of the different ways antibodies can be used as therapeutics. **a** antibodies which are used as TCR-mimics; **b** Immunoglobulin antibodies and different antibody fragments which are used as therapeutics; **c** antibody-drug conjugates (ADCs) and how they should trigger apoptosis; **d** antibodies which target a malignant cell and then recruit immunogenic cells (T cells, macrophages, etc.) - multifunctional or bispecific antibodies.

- Long half-life: Antibodies have a longer half-life and thus a better bioavailability in the body compared to small molecules, allowing for less frequent dosing and improved patient convenience [29][30].
- Diverse applications: Antibodies have been used as therapeutic drugs for the treatment of various human diseases, including cancer, asthma, arthritis, psoriasis, Crohn's disease, transplant rejection, migraine headaches, and infectious diseases. They can be engineered to enhance their safety and efficacy, and new antibody-based drugs are continually being developed [25][31].

The benefits of antibody-based therapies have revolutionized the treatment of many diseases, and their use as drugs continues to grow, with significant potential for future advancements. Being able to specifically engineer antibodies with a desired function is therefore of great interest and will be able to combine the benefits of *de novo* protein design with the clinical relevance and potential of antibodies as therapeutics.

The authors of [32] could show that protein binder design is working successfully for a lot of previously unsolved challenges. However, in order to successfully validate the designed antibodies - analogous to the binder design as shown in [33] - general antibody structure prediction will need to be improved. As already mentioned, there are several aspects to antibodies which make this a hard problem. Antibodies exhibit a significantly high level of structural and sequence-based similarity and only a small part of the antibody is involved in the interface where the antigen is bound. This part of the antibody - also called the complementarity-determining regions (CDRs) or hypervariable region, lacks a distinct secondary structure, making it exceedingly challenging to model using current state-of-the-art structure prediction methodologies such as AF2 or RF2 [21]. For a better visual representation an entire structure of an antibody with the CDRS highlighted can be observed in Figure 1.2.

1.4 Improving Structure Prediction with a sequence-only Data Set

Assuming that the architecture of either RF2 or AF2 is ideal for antibody structure prediction, the underlying issue could be primarily attributed to a lack of data. Consequently, leveraging the vast sequence space to enhance antibody structure prediction represents a promising avenue for exploration. Notably, the Observed Antibody Space (OAS) data set contains over two billion antibody heavy chain sequences and is continually expanding[35]. This extensive data set holds the potential to significantly augment the data available for improving the accuracy of antibody structure predictions.
The motivation behind this approach is rooted in AlphaFold2's self-distillation technique, as detailed in [9]. An integral component of this improvement lies in the capacity of AlphaFold2 to provide a per-residue metric for each prediction, known as the per-residue local distance difference test (pLDDT).
For the self-distillation technique, AF2 predicts structures for a subset of the Uniclust30 database [36] which is a database of clustered protein sequences and consists of around 6 million sequences. This particular subset encompasses around 350,000 sequences. Subsequently, after

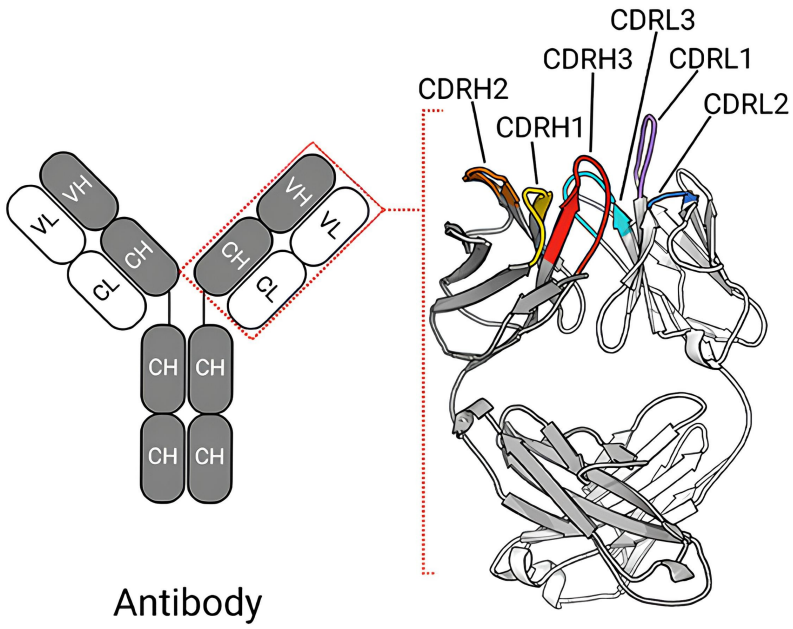


Figure 1.2: Figure taken and modified from [34]

predicting the structures of these sequences and filtering them based on a pLDDT-like metric, this subset was utilized for retraining AF2 in conjunction with PDB-structures. This approach was able to improve the overall structure prediction accuracy of AF2 by a significant margin. Figure 1.3 shows a schematic of the described self-distillation technique.
Furthermore, it is essential to note that not only AlphaFold2 but also RF2 underwent training using a distillation approach, leading to a clear improvement in the overall structure prediction accuracy of RF2, as outlined in [18].

1.5 Aim of this Thesis

The aim of this thesis was to achieve two objectives: the first was to demonstrate that fine-tuning RFantibody on a sequence-only data set could improve the prediction of binding probabilities for antibody-antigen complexes demonstrating that the model can reason using the newly gained understanding of structures, whereas the second was to showcase that this fine-tuning process could enhance the overall accuracy of antibody structure prediction. It has to be pointed out that the prediction of binding probabilities represented the more straightforward of the two tasks, and is a prerequisite for the second task. Consequently, the first objective was pursued first, being the first hallmark of successfully incorporating sequence-only data.

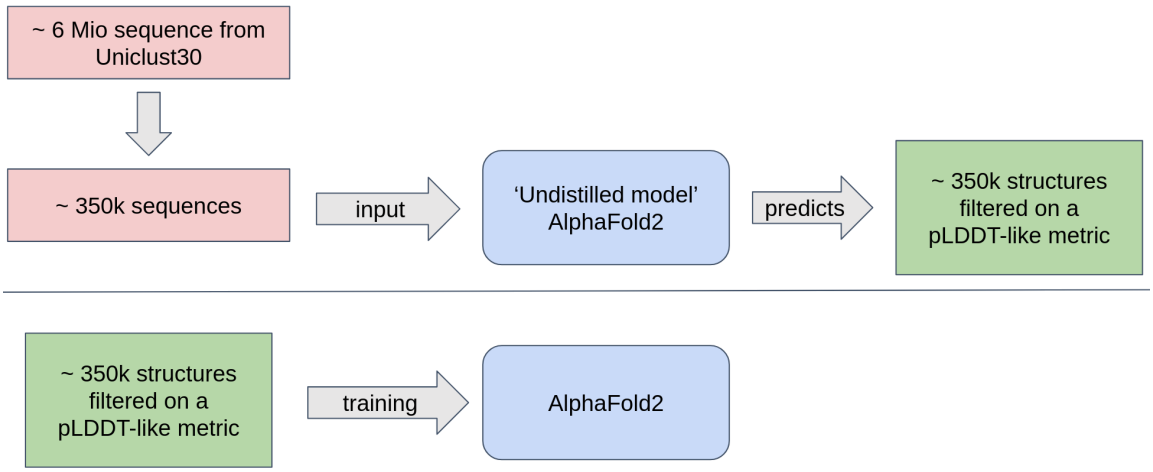


Figure 1.3: The pipeline of how AF2's self-distillation approach works. AF2 is trained on a subset of the Uniclust30 database and then the predicted structures are filtered based on a pLDDT-like metric. The filtered structures are then used to further train AF2.

2 Methods

2.1 RoseTTAfold2's Architecture

This section is by no means a complete explanation of the architecture of RF2 but rather a brief overview of the most important parts of the architecture. For a more detailed explanation refer to the original paper [18].
The input to RF2 consists of three different types, each fed into the three separate tracks of the architecture as outlined in Figure 2.1. The *1D track* incorporates the MSA of the sequence of interest.
The *2D track* or pair-wise representation takes a 2D representation of the 3D structure derived from the templates (when available). Templates represent homologous structures to the sequence of interest and are generated using the hhsuite software suite [37].
The *3D track* encompasses the 3D coordinates, which are subsequently channeled into a SE(3)-equivariant transformer, as depicted in the schematic in Figure 2.1. A particular type of self-attention is used to fulfill the criteria of SE(3)-equivariance.
The *1D track* feeds the *2D track*, where a pairwise representation of the prediction develops. This representation feeds back to the *1D track*, such that attention aids processing the MSA. Both tracks feed the *3D track*, where a 3D structure is developing. This 3D structure representation also feeds back to the other tracks refining the other representations.
The integration of these three tracks enables synchronization, ensuring that the sequence, pair-wise representation, and 3D structure remain in correspondence throughout the one pass-through through the network. One of the features which greatly improved structure prediction was recycling the predicted structure and the embeddings for the pairwise representation and for the MSA as an input for the next iteration - therefore giving each track its own recycling input. The refinement block as depicted in Figure 2.1 is intended to refine the predicted structure which is why it is only computing attention from each residue to its 64 nearest neighbors.
In addition to producing the final structure, the network also generates the metric *pAE* (predicted aligned error), which serves as a pairwise metric estimating the error of the predicted structure. Furthermore, it yields *pLDDT* as previously described, and *pBind* - a probability that two chains are binding computed from *pAE* logits. The model is capable of doing so, because during training it has to reason whether or not two provided proteins will form a complex or not using its structural understanding. However, the training data set also contains *negative* examples in the form of generated complexes with a *pBind* of 0.0 assigned to them.
The network is trained on a data set consisting of 280k structures from the PDB along with an AF2 generated distillation set which alone consists of 3.6M sequence/structure pairs. The loss function incorporates terms like distogram loss, masked amino acid prediction loss, structure loss, pAE and pLDDT prediction losses, bond geometry loss, clash loss, and predicted binding loss. The training occurs in two distinct phases, with specific parameter settings, such as

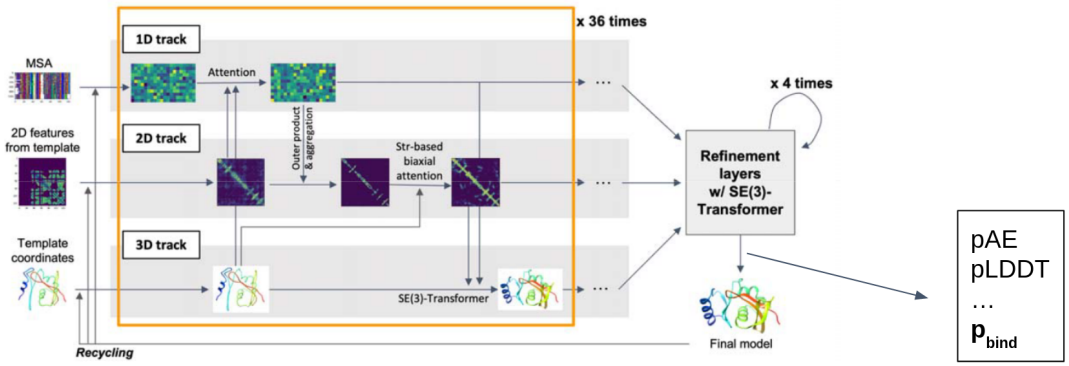


Figure 2.1: A schematic showing the architecture of RoseTTAfold2 where each track and its respective inputs are shown: **1D track** - consisting of the MSA of the sequence of interest; **2D track** - consisting of the contact map of the templates; **3D track** - consisting of the 3D coordinates of the templates. These parallel tracks update the different representations and are then fed into a refinement block which then generates the final structure which is then recycled as an input for the next iteration. Moreover, this architecture yields additional features such as predicted aligned error (pAE), per-residue local distance difference test (pLDDT), and *pBind*. Of particular significance for this thesis is the *pBind* feature. (Figure taken and modified from [18])

loss weights, crop size, and learning rate, tailored for each phase. Moreover, the approach incorporates cropping techniques for large proteins to optimize the training process.

2.2 Hotspots for RFantibody

Unlike RF2 which aims to solve general structure prediction because this alone is a complicated task, RFantibody is specifically tailored to predict the structure of antibodies as finding and correctly predicting the antibody-antigen dock is particularly hard. To achieve this objective, the authors of (manuscript in preparation) discovered that providing privileged information - here termed **hotspots** - about the binding interface of an antibody to RFantibody significantly enhances the overall structure prediction accuracy as depicted in Figure 2.2.
These hotspots are defined as residues on the target structure with an average C_β distance from $10\,\text{Å}$ to any residues on CDR loop on the antibody. These identified hotspots are subsequently incorporated as an additional input layer within the *1D track* of the network therefore being assigned as a per-residue feature. One of the primary challenges addressed in this thesis revolved around attempting to replicate this privileged information within this specific sequence-only data set.

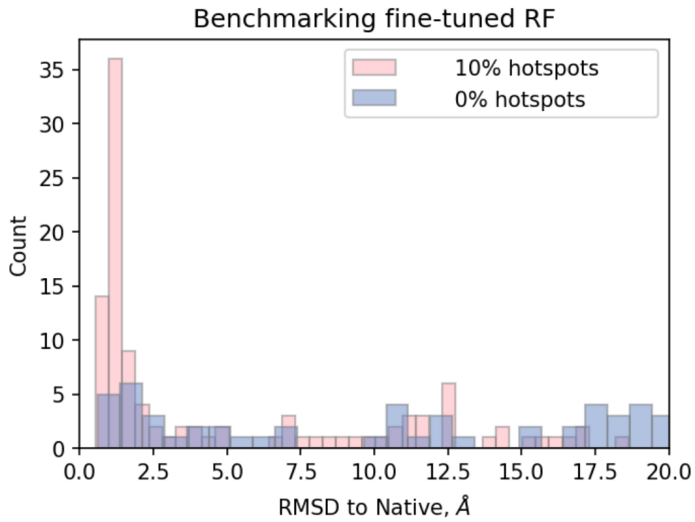


Figure 2.2: This graph shows the improvement in structure prediction accuracy when giving RFantibody either no hotspots or 10 % hotspots. The y-axis shows the RMSD in Å to the native structure and the x-axis shows the number of structures within that RMSD bin.

2.3 Data Set Preparation

The data set utilized in this study was sourced from Cao et al. [38], forming the cornerstone of this thesis. The entire workflow was written with Python 3.10 [39]. Data processing and manipulation was done using the pandas library [40], an especially useful datatype was the pandas dataframe for its fast access. The clustering process was done using the subprocess library from Python to call *MMseqs2*. Other useful libraries were numpy [41] for processing and matplotlib [42] and seaborn [43] were used for data visualization. The PyTorch library was used to generate the necessary data structures for training [44].

2.4 Benchmarking

We set up a benchmarking test to measure performance on the held-out set, with a primary focus on evaluating the predictive performance of the *pBind* metric. Alongside that, the *pLDDT* and *pAE* metrics were also closely monitored during the evaluation process. Furthermore, the RMSD to the native structure was also calculated which was only possible for those antibodies within the data set that had an associated structure, a criterion that was met by a mere 76 out of the 3051 antibodies. Initially, the held-out set was selected using the *t*-SNE epitope clustering method, but subsequently, the sequence similarities of the concatenated CDRH3 and CDRL3 were used to have a different metric of clustering to compare to the already established clustering.
For the analysis - as shown in subsubsection 3.2.3 the metrics *pBind* and *pAE* were plotted. Each metric was accompanied by a density plot illustrating the distribution of positive and negative samples. Moreover, the reciever operating charateristic (ROC) area under the curve

(AUC) was calculated for only a specific subset of the held-out set, i.e. only for predictions where the *pAE* was smaller than 10 or 5.

2.5 Training

As already mentioned in section 2.1 RFantibody is fine-tuned off of the published weights of RF2. This process involved an initial training phase on antibody-specific structures, followed by a subsequent sequence-only fine-tuning step. However, the ratio of sequence-only samples was never more than 50% of the total training samples to avoid overfitting on that respective data set and to avoid the model collapsing in on the task of general structure prediction. This also implies that the total number of samples the model sees from the structure based training set increases during the trainings conducted for this thesis. This notably can have side-effects such as overtraining or a general improvement in overall structure prediction not attributed to the sequence-only data set.

Training procedures were executed using 8 NVIDIA L40 GPUs, with a batch size of 64 and the AdamW optimizer [45]. If no improvement of the validation loss for the *pBind* metric was observed within a 2-week period, the training process was halted. In hindsight, it is clear that the extensive training period was far beyond what was necessary for the models that were not effectively learning, reflecting a now better understanding of neural networks. These training operations were conducted on a local high-performance cluster (HPC), while monitoring was facilitated using the Weights and Biases framework [46].

2.6 Loss Functions

In the context of RFantibody, taking over the exact same loss terms like RF2 [47], the similarity between the loss terms used in structure prediction training, involving known structures, is highlighted. However, a notable challenge arises during the training scenario focused on sequence-only data, where the absence of explicit structures limits the applicability of several loss functions. Each loss also has a weight term associated which determines the magnitude of how much it will be considered and is determined during a hyperparameter search. The following loss terms are usually applied during training with structures:

- **Distogram loss** - Cross-entropy between the predicted pair-wise representation and the native.
- **Structure loss** - Consisting of the frame-aligned point error FAPE term, a specific chirality-aware representation, where each residue is represented as a triangular frame (spanning over C_α-C_β-N)
- **Loss on the networks' pAE prediction** - Cross-entropy between predicted and true pAE.
- **Loss on the networks' pLDDT prediction** - Cross-entropy between predicted and true pLDDT.

- **Loss on bond geometry** - Sum of losses on peptide bond deviation from ideal, with normalized length and angle terms.
- **Clash loss** - Uses Rosetta's Lennard-Jones potential, normalized by the number of residues, providing weak attraction and strong repulsion between atoms.

For the sequence-only approach pursued in this thesis, the following loss functions can be applied:

- **Masked amino acid prediction loss** - Cross-entropy between actual and predicted sequences over randomly masked positions.
- **Predicted binding loss** - Binary cross-entropy between the *pBind* head given that a pair of chains forms a complex.

Loss terms, like Distogram loss, Structure loss (comprising frame-aligned point error FAPE, chirality representation), pAE and pLDDT prediction losses, bond geometry loss, and clash loss, are conventionally applied when structural information is available. In contrast, for the sequence-only approach pursued in this study, specific loss functions come into play, including Masked Amino Acid Prediction loss and Predicted Binding loss. The inherent challenge lies in the model's diminished ability to learn when confronted with a reduced set of applicable loss functions, emphasizing the importance of specific strategies for training in the absence of explicit structural information. By adding structural examples the model should be incentivized to always reason using the structural representation.

3 Results and Discussion

The subsequent chapter focuses on presenting the challenges which arose during the data set preparation and the extractions of the privileged information leading to a selected number of training runs and associated benchmarking results. It is important to note that many other training runs produced similar outcomes and this only reflects a small subset. The training runs which are presented here are the ones which show the highest diversity in response to the different training settings.

3.1 Curation of a sequence-only data set for RFantibody

The data set used for fine-tuning RFantibody on a sequence-only data set had to fulfill certain criteria. Notably, it had to be a large enough data set and it had to be associated with a specific target. For this purpose, the SARS-CoV-2 spike protein was selected as the target due to its extensive research focus, resulting in a substantial availability of associated antibody sequences. Additionally, there needed to be some additional information in order to generate hotspots (as explained in section 2.2) on the target, which were proven to be essential - especially since there were no structures associated.
The data set [48] used, encompassed 3051 monoclonal antibodies (mAbs) accompanied by binding affinity data linked to every single point mutation on the target protein. This data was obtained using deep mutational scanning (DMS) - on the target, a 194 residue-long receptor binding domain (RBD) of the SARS-CoV-2 spike protein as depicted in the schematic Figure 3.1. Therefore the total number of data points was around 10M, making it a very promising data set. The resulting binding values were normalized and then subsequently utilized to compute escape scores ranging between 0 and 1 using an internally developed equation [38]. These scores were subsequently correlated with the *pBind* metric of RFantibody. This information was used for generating the hotspots on the RBD as elaborated in section 2.2.

3.1.1 Clustering

The authors of [48] applied t-distributed stochastic neighbor embedding (t-SNE) [49] to cluster the 3051 antibodies along with their DMS profiles into 12 different clusters (or epitope groups) as demonstrated in Figure 3.2. The t-SNE algorithm serves as a non-linear dimensionality reduction technique, facilitating the visualization of high-dimensional data within a low-dimensional space.
Initially, these clusters were used for the train/validation split but in order to see if binding and then structure prediction can be improved when the clustering is performed in sequence space, the *easy-linclust* method from the software suite *MMseqs2* was utilized [50]. Notably, this approach solely utilized the concatenated sequences of the complementarity-determining region of the heavy chain (CDRH3) and the complementarity-determining region of the light

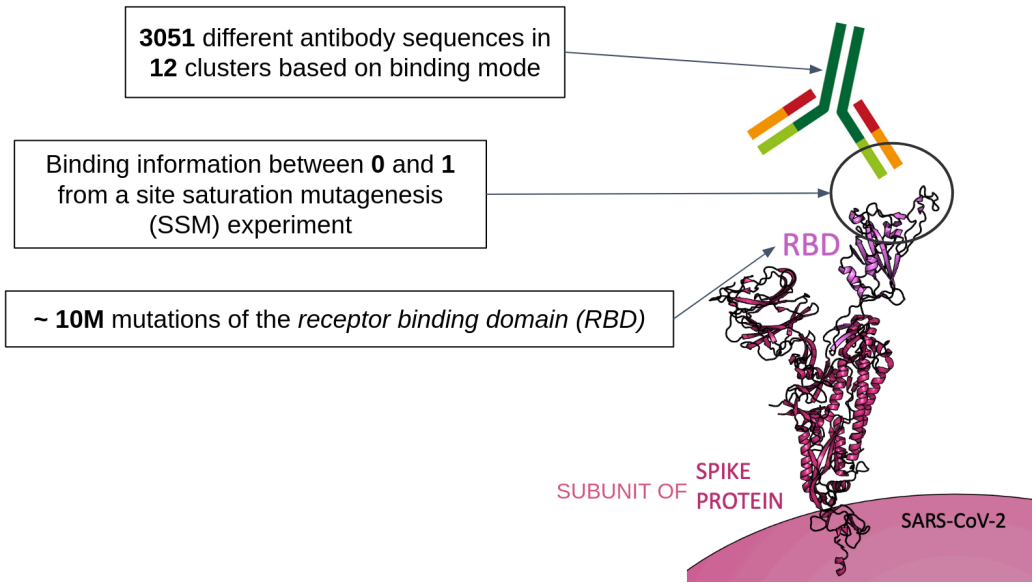


Figure 3.1: A schematic showcasing the data set used. 3051 antibodies in 12 different cluster modes associated with 10M mutations of the RBD - a subunit of the spike protein of the SARS-CoV-2. Binding data is derived from a site saturation mutagenesis (SSM)-experiment [48]. Figure taken and modified from [34].

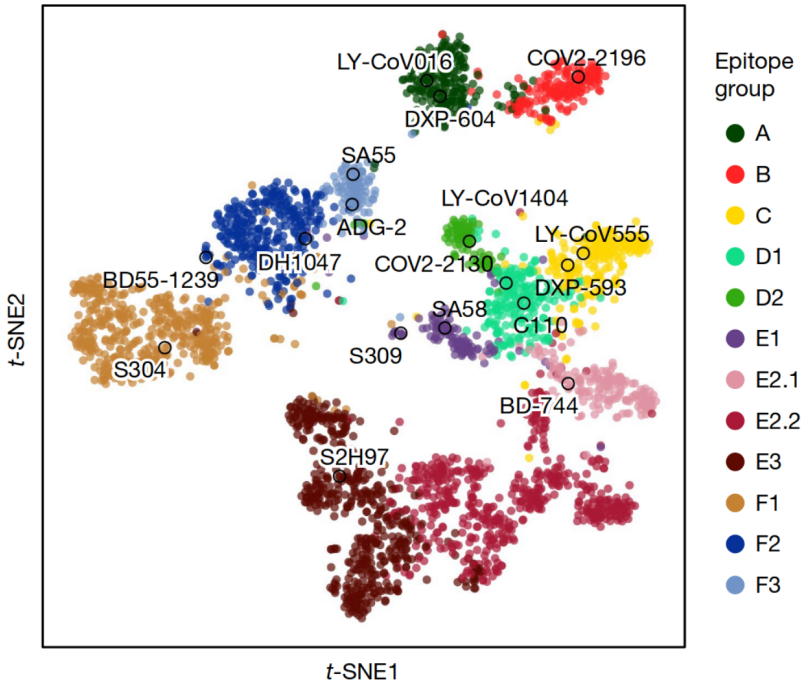


Figure 3.2: The resulting epitope clusters after performing *t*-SNE on the DMS profiles of 3051 antibodies performed by [48].

chain (CDRL3), as illustrated in Figure 1.2. The utilization of this concatenated sequence data yielded the generation of the fewest clusters, a feature highlighted in Figure 3.3. As a result of this outcome, the combined CDRL3 and CDH3 were used to perform the subsequent clustering in sequence space and thus the train/validation split to compare it with the other clustering approach.

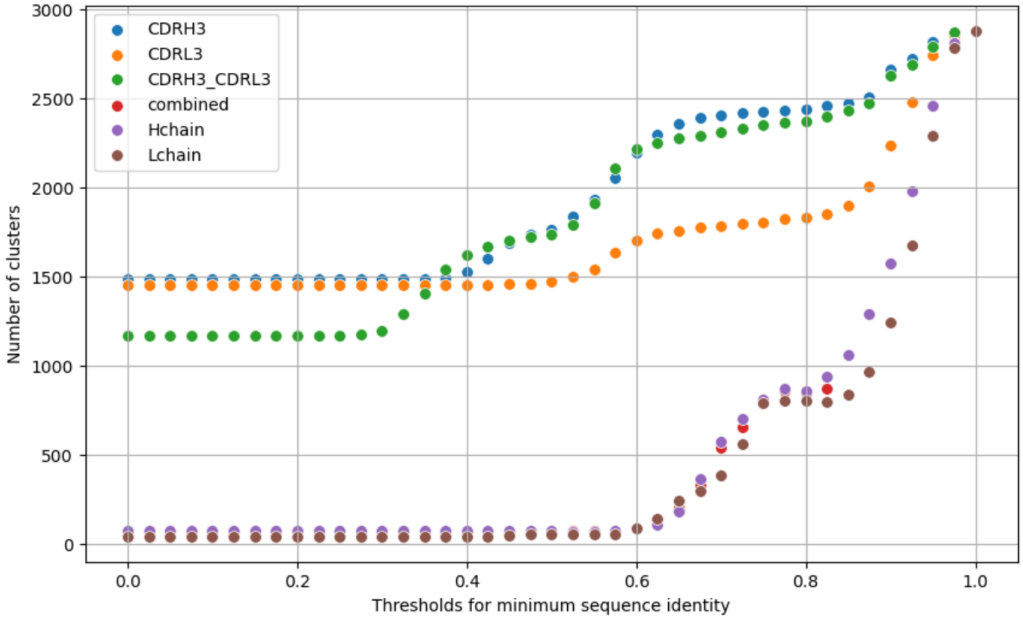


Figure 3.3: After using the *easy-linclust* method from the software suite *MMseqs2* [50] on the data set of 3051 antibodies the least amount of clusters was achieved when using only the concatenated sequences of CDRH3 and CDRL3 as the green points suggest. The clustering was performed in 0.25 similarity increments.

3.1.2 Hotspot Selection

Here, each antibody in conjunction with each mutation on one of the 194 residues on the RBD, has an escape score mapped between 0 and 1 associated with it. To establish the definition of a hotspot, an arbitrary threshold was selected. Depending on the score and the threshold a residue on the RBD is either declared as hotspot or not. One additional criterium was that the residue had to be on the surface of the RBD, determined through the utilization of the Rosetta software suite [51], using the same structure for the RBD as the curators of the data set [48] with the PDB identifier 6M0J [52]. This information was compiled as a boolean mask with the length of the RBD serving as additional data associated with each antibody, as depicted in Figure 3.4.

3.1.3 Positive Negative Split

The data set had a bias, with a significantly skewed ratio between positives (i.e., antibody-RBD interactions which are binding) and negatives (i.e., antibody-RBD interactions which are non-

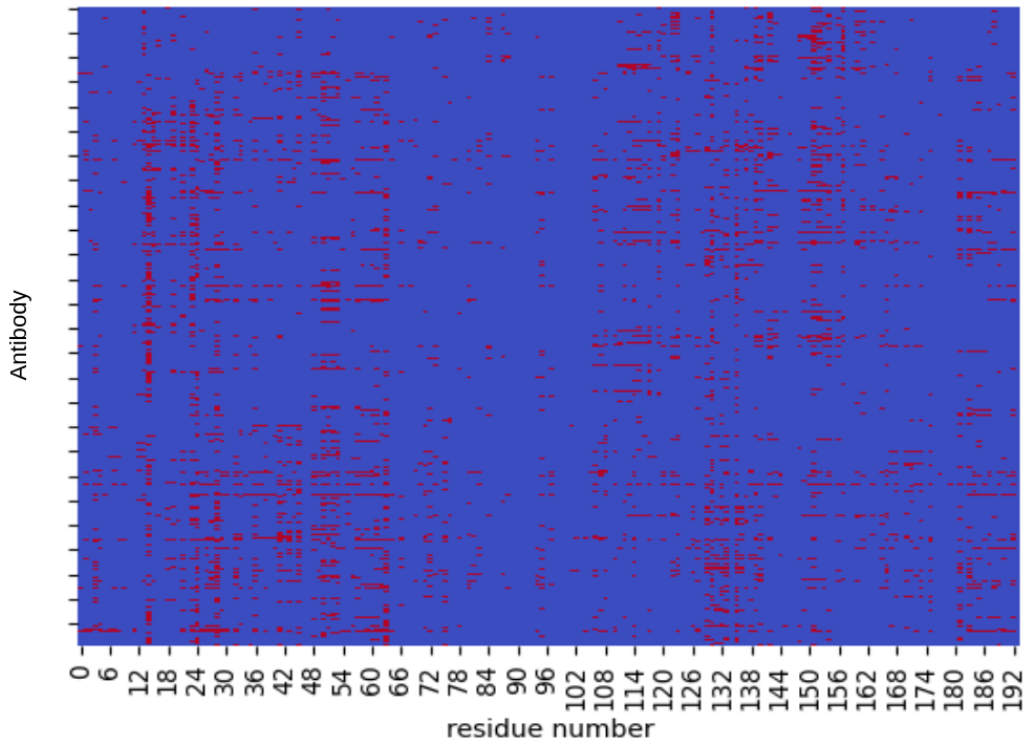


Figure 3.4: A heatmap showing the hotspots on the RBD of SARS-CoV-2. The x-axis shows the 194 residues of the RBD and the y-axis shows the 3051 antibodies. The color red indicates that a residue is a hotspot and the color blue the opposite.

binding). The inbalance was due to a stark over-representation of binding interactions. This can be observed in the violin plot in Figure 3.5. Ideally, the ratio of positive to negative samples should be maintained at a close balance of 1:1. In most cases, failure to achieve this balance often results in the model consistently predicting either only binding or only non-binding events, depending on the predominance of one class over the other in the training data set.
To address the issue of imbalance, an arbitrary cut-off (see green line in Figure 3.5) was established above which a sample was considered a negative or non-binding event. Subsequently, an equal amount of positive or binding samples was then chosen from the highest value of the positive binding events. This imbalance correction presented a significant challenge as it led to a drastic reduction in the overall number of samples available for training.

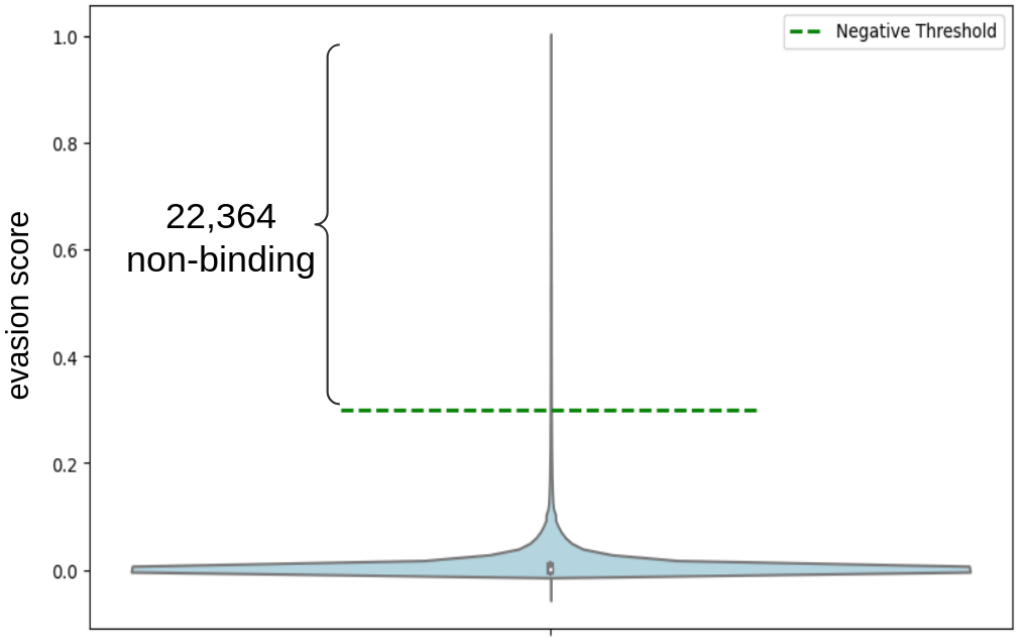


Figure 3.5: A violin plot depicting the imbalance of positive and negative binding events. The y-axis shows the evasion score meaning that 1 correlates with non-binding and a negative example whereas 0 correlates with a positive binding event. The green line shows an arbitrary cut-off above which a sample is considered a negative or non-binding event. For that specific value of 0.3 the data set is reduced to in total around 44k samples.

3.1.4 Bias Elimination

In the course of training of the model and the analysis of the benchmark results it became evident that there was another bias in the data set. This introduced bias was primarily a result of an initial lack of thorough investigation into the data set on my part. The bias was caused by the fact that most non-binding or evasion events occurred predominantly through mutations at specific residue positions.

This can be observed in Figure 3.6 where the x-axis shows the 194 residues of the RBD and the y-axis shows the frequency of occurrence of a residue in the data set. For instance as depicted the residue on position 52 very often results in a negative binding event (i.e. evasion), whereas the residue on position 51 rarely resulted in such events. A similar pattern is observed in the data set containing positive binding events (i.e., successful antibody-virus inhibition), however it is not as pronounced as for the negative binding events. Consequently, the model could learn this position-specific pattern rather than reasoning over the structure of the binding interface. We want the model to be learning whether a specific mutation would affect binding to a specific antibody, but this bias means that the model can simply learn that specific residues positions (irrespective of antibody sequence) are associated with binding/non-binding. This issue was further exacerbated when one of these critical residue positions was present in both the training and validation sets, ultimately resulting in the model learning this information. This can be observed in the reciever-operating-characteristic (ROC) Figure 3.10 which shows clear indications of the model having found this shortcut.

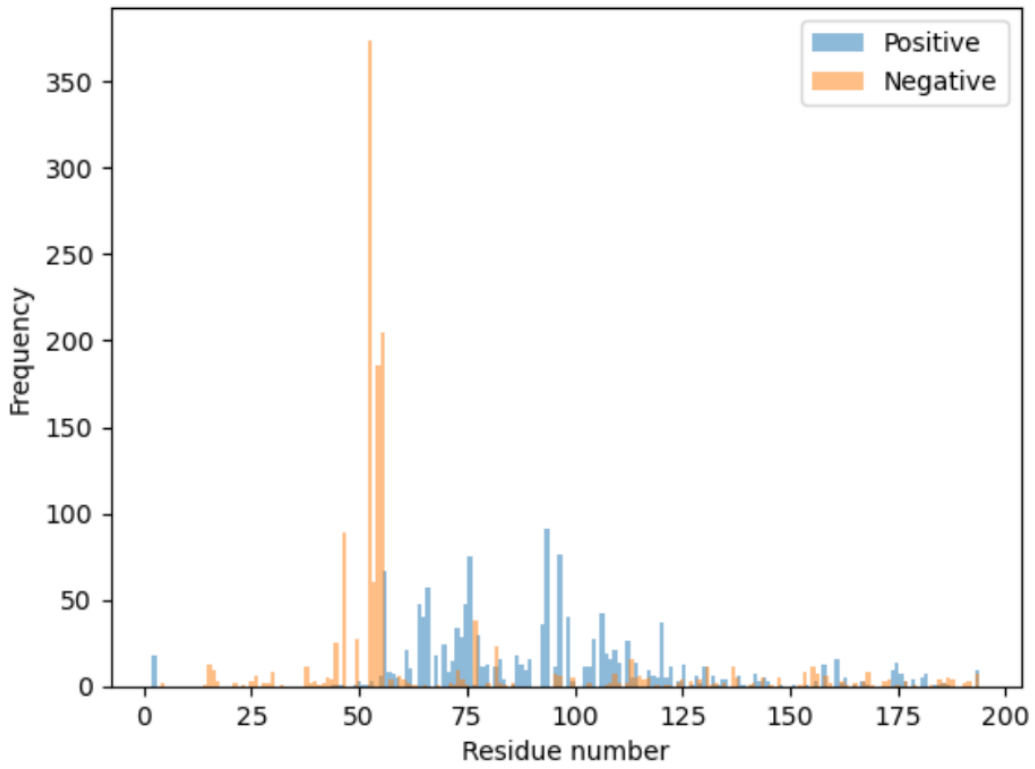


Figure 3.6: A bar plot showing the frequency of occurrence of a residue in the data set. The x-axis shows the 194 residues of the RBD and the y-axis shows the frequency of occurrence of a residue in the data set. The color orange is used for the non-binding events and the color blue for the binding events.

3.2 Evaluation of individual Models

The following chapter contains the different iterations of preprocessing (e.g. clustering) and subsequent training approaches employing different hypotheses which were tested if they could improve the binding prediction as a first step towards structure prediction.

3.2.1 Model without any fine-tuning

These benchmarking results aim to provide a baseline comparison since this model was not fine-tuned with any sequence-only data set but instead the exact same benchmarks were performed on the weights from RFantibody which were used to fine-tune the other models from below.
As depicted in Figure 3.7 and Figure 3.8 the not fine-tuned version of RFantibody exhibits an acceptable behavior for the distinction between binding and non-binding events. However, the ROC AUC of 0.63 for *pBind* and 0.67 for *pAE* - though being very impressive - can certainly be improved. However, for this thesis it can be regarded as the gold standard in comparison to the following training runs. The density plots demonstrate a comparable pattern.

3.2.2 Model with suspected Data Leak

For this case neither the different clustering approach (as explained in Figure 3.3) nor the bias elimination (as explained in Figure 3.6 and Figure 3.5) were employed since the issues elaborated in subsection 3.1.4 and subsection 3.1.3 were not yet considered. Here, the loss curves for the *pBind* metric show indications of strange behavior (potentially indicating issues with the training set) as depicted in Figure 3.9. This became more evident during the analysis of the benchmarking results as discussed in the following sections.

Training

The model was trained on the *t*-SNE epitope clusters and only after the model had seen around 100k samples did the loss for the *pBind* metric begin to decline. This progression is illustrated in Figure 3.9, where the x-axis represents the number of steps and the y-axis denotes the loss for the *pBind* metric.

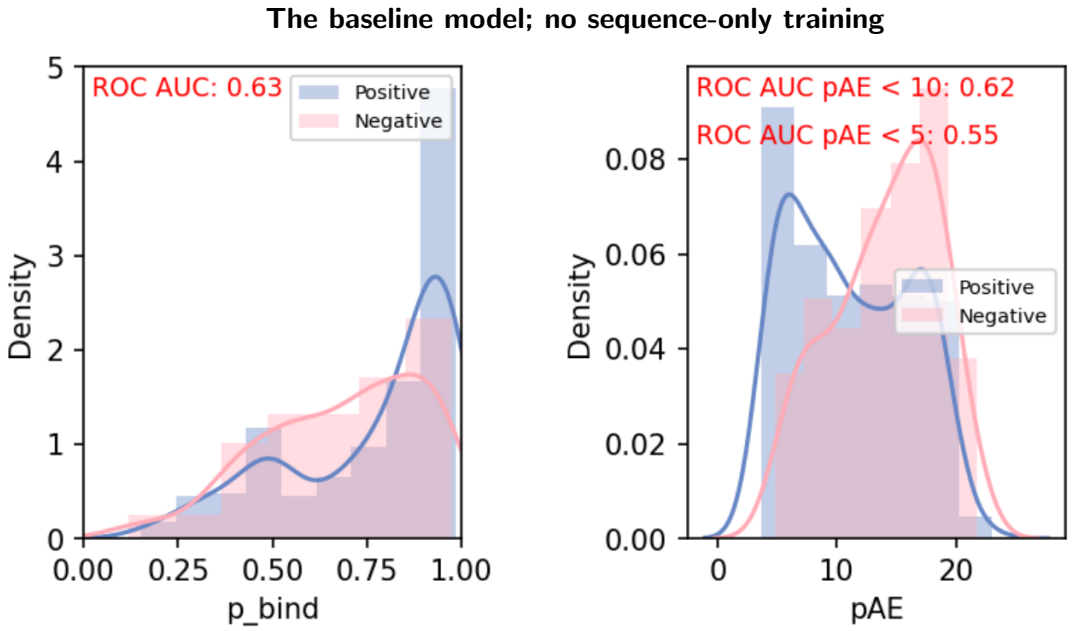
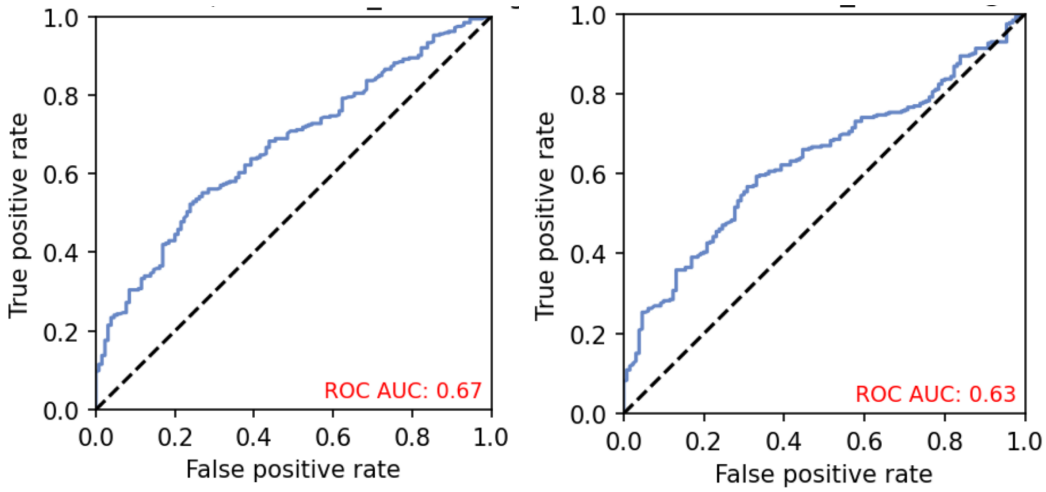


Figure 3.7: Distograms for two metrics: *pBind* and *pAE*



(a) The ROC curve for the *pAE* metric.

(b) The ROC curve for the *pBind* metric.

Figure 3.8: RFantibody without any fine-tuning was evaluated to have a baseline model to compare the models against with AUCs for *pBind* and *pAE* being 0.63 and 0.67 respectively.

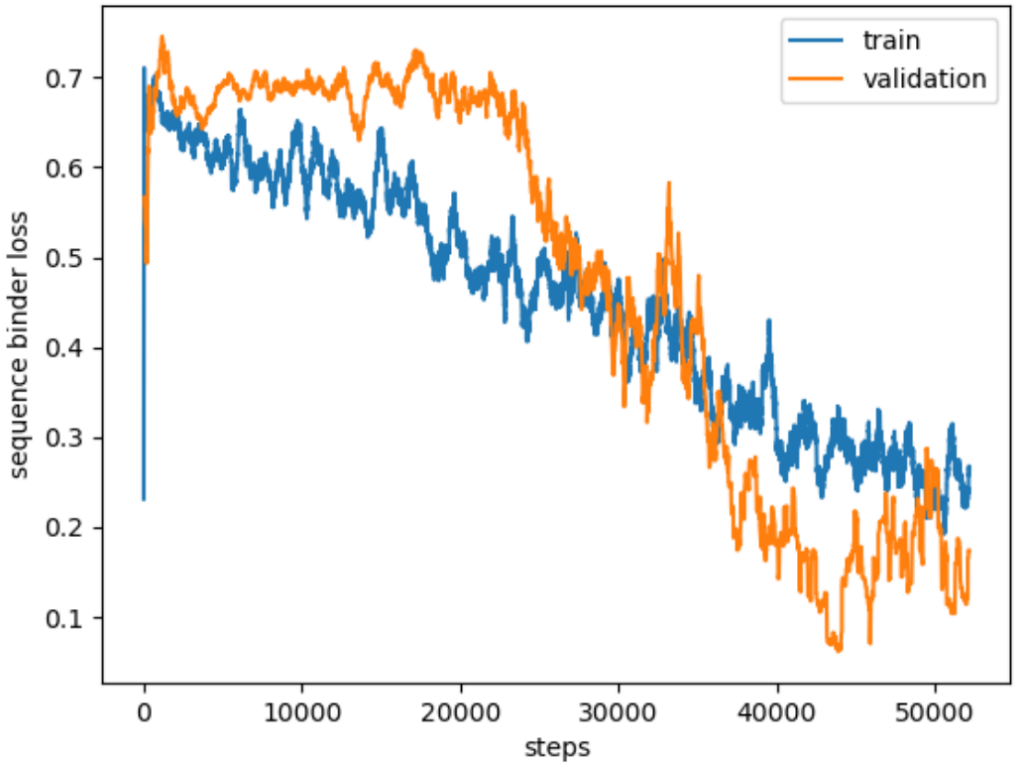


Figure 3.9: The orange line represents the validation loss and the blue line represents the training loss. Here depicted is the binding loss for the *pBind* metric for the model with suspected data leak and unexpected behavior. The x-axis shows the number of steps and the y-axis shows the binding loss for the *pBind* metric. Since the validation loss (orange line) gets smaller than the training loss (blue line) the model is not just overfitting on the training data but also learning something from the training data to predict the validation data beyond the losses applied - a data leak could be hypothesized. The overall trend of the losses also exhibits unusual behavior.

In Figure 3.9 the orange validation loss stagnates, while the training loss very slowly decreases, which is a sign of potential underfitting - a phenomenon which occurs when the complexity of the problem is too high for the model to extrapolate from the training set to the validation set. However, at approximately 22k steps the validation loss begins to decrease and after around 30k steps the validation loss is lower than the training loss. The fact that the validation loss starts to decline abruptly can be an indicator of issues with the training set. Taking both loss curves into consideration this training run indicates a lot of erroneous conduct and close attention was paid to further analysis of the data set.

Benchmarks

The hypothesis of a data leak comes when looking at the results of the benchmarking. The benchmarking was conducted on the held-out set selected using the *t*-SNE epitope clusters. The benchmarking results depicted in Figure 3.10 show the ROC-curve. The data leak becomes apparent when the graph indicates that at a true positive rate of 80% the false positive rate is 0%. This scenario implies that 80% of the binding events are correctly identified as such, and all non-binding events are correctly identified as negative, indicating a perfect performance. Furthermore, the AUC metric shows a value of 0.96 which is very close to the ideal value of 1.0 - this is a significantly better performance than expected. Taking the loss curves from Figure 3.9 into consideration the potential of a data leak is substantiated. The subsequent investigations revealed the biases which yielded a potential explanation for the loss curves and the benchmarking result - those biases were eliminated as explained in Figure 3.6.

3.2.3 Model without Bias

Working with the data set, that was normalized with respect to the residue position. The normalization of the data set however greatly decreased the size of the data set. Additionally, the hypothesis was that clustering in sequence space would improve the model. To test this, a new clustering was performed using the *easy-linclust* method from the software suite *MMseqs2* [50]. Subsequently, the model was retrained on the modified data set. The results of that training and benchmarking can be observed in the following sections.

Training

The loss curves for the *pBind* metric can be observed in Figure 3.11. It should however be noted that over the whole training process, the validation loss only exhibited a very small downward trend, suggesting that the model struggled to learn from the training set. Considering the reduced size of the data set due to normalization, the model was limited to observing 25k training samples before the training was terminated.

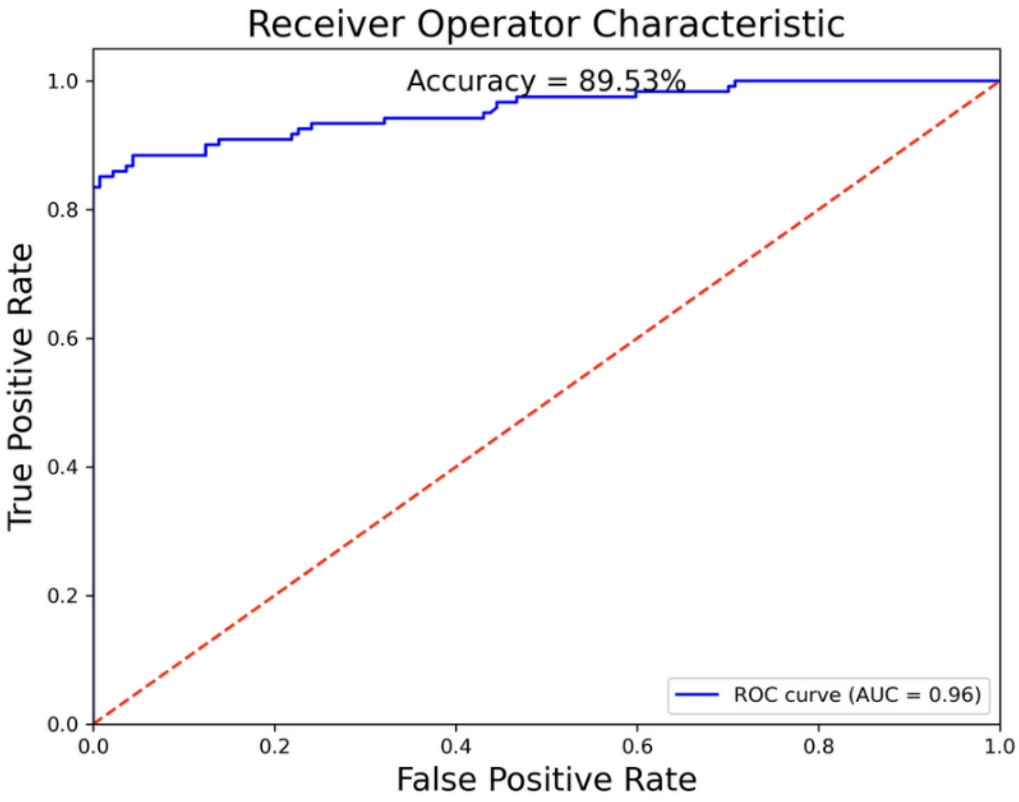


Figure 3.10: The ROC graph for the benchmarking results of the data set with suspected data leak. The x-axis shows the false positive rate and the y-axis shows the true positive rate.

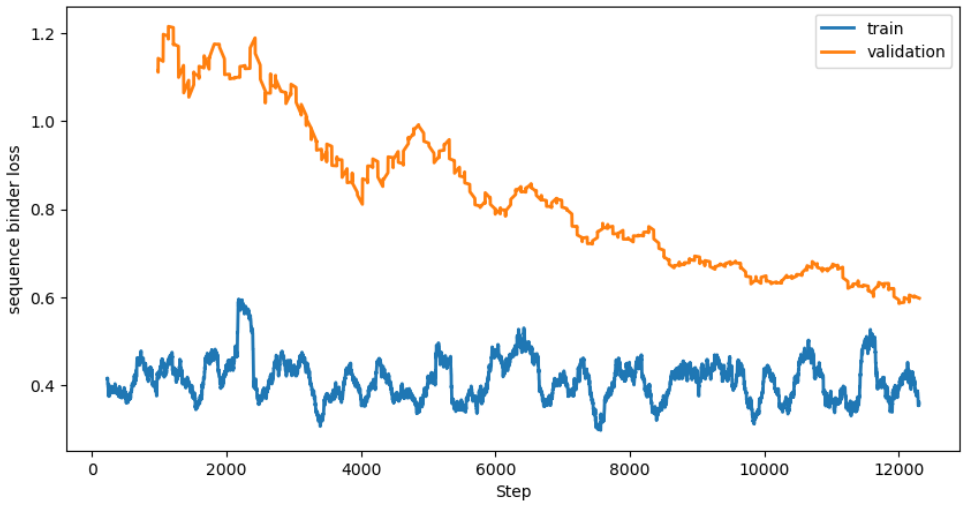


Figure 3.11: The binding loss for the *pBind* metric. The x-axis shows the number of steps and the y-axis shows the binding loss for the *pBind* metric. Note here that the models in subsection 3.2.4,subsection 3.2.5, and subsection 3.2.6 show an almost identical behavior for their individual loss functions which is why they are not shown here.

Benchmarks

Additionally, to the bias elimination and the new clustering approach the benchmarking was also conducted more extensively, encompassing the *pBind* and *pAE* metrics with the latter being also used to distinguish between binding and non-binding events.
The metric *pBind* is computed using the 64 binned *pAE*. However as depicted in Figure 3.12, *pAE* is more suitable for distinguishing binding events. Notably, both metrics do not perform well, considering that an ROC AUC of 0.62 does not hold great discriminative power. The density plot for *pBind* very clearly shows that the model is not able to learn the task since the distribution of the positive and negative binding events is almost identical. This indicates that it was not able to generalize for negative binding events and instead predicts them as positive instead. The ROC curve in Figure 3.13 clearly reflects the same trend, where a slightly better than random discrimination is achieved. If anything, it can be stated that the overall performance in comparison to the model without any fine-tuning is slightly worse.

3.2.4 Model with different Hotspot Threshold

This model was trained on the same data set as the model in subsubsection 3.2.3 but with a different threshold for the hotspots. The leading hypothesis for choosing a different threshold for the hotspot was that a lower threshold would incorporate a higher amount of residues on the target resulting in a higher diversity and giving the model more options for predicting the dock correctly. The threshold was set to 0.5 instead of 0.3 above which a value was considered a mutation causing evasion. The results of the training and benchmarking can be observed in Figure 3.15 and Figure 3.14. However, upon comparison to the baseline model, no significant difference was observed, which is why this model was not further investigated.

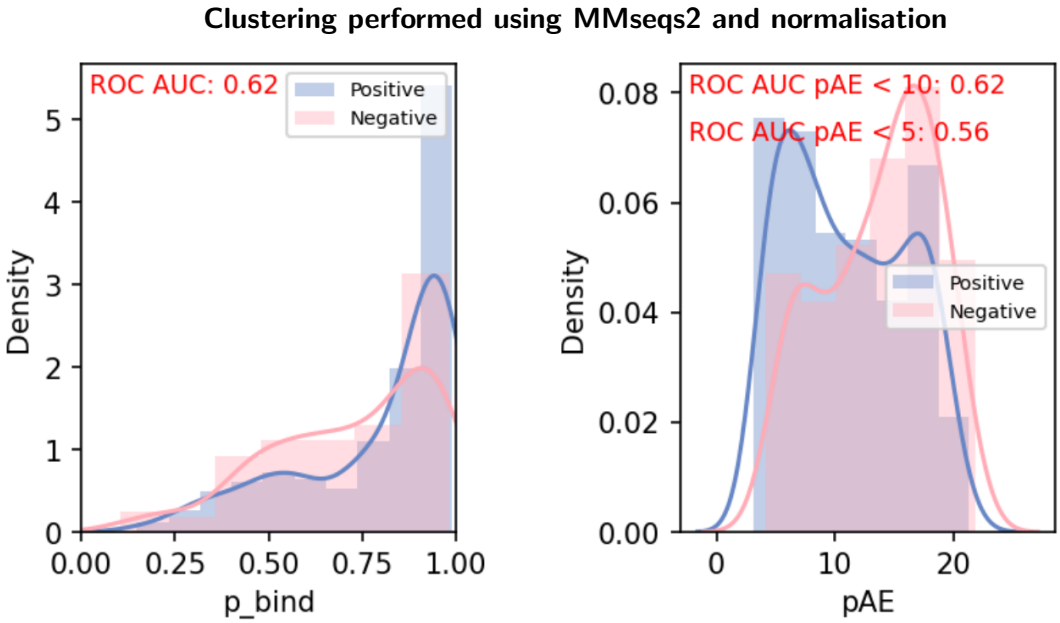
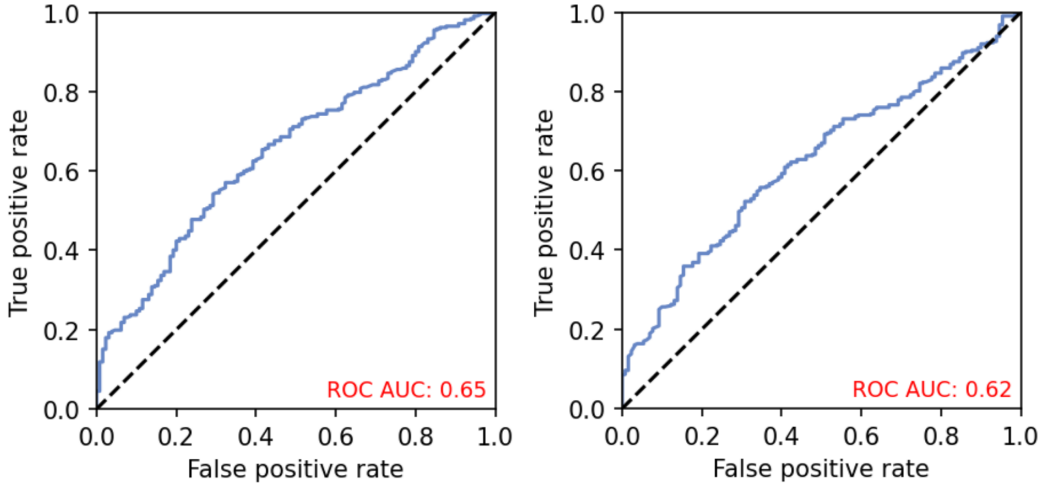


Figure 3.12: Distograms for two metrics: *pBind* and *pAE*



(a) The ROC curve for the *pAE* metric.

(b) The ROC curve for the *pBind* metric.

Figure 3.13: RFantibody was evaluated on two metrics; *pBind* and *pAE*. Changing clustering from epitope-based to sequence-based and eliminating the residue-specific bias during training only led to a marginal decline in the distribution of metrics between the binding and non-binding set, with AUCs for *pBind* and *pAE* being 0.62 and 0.65 respectively.

Benchmarks

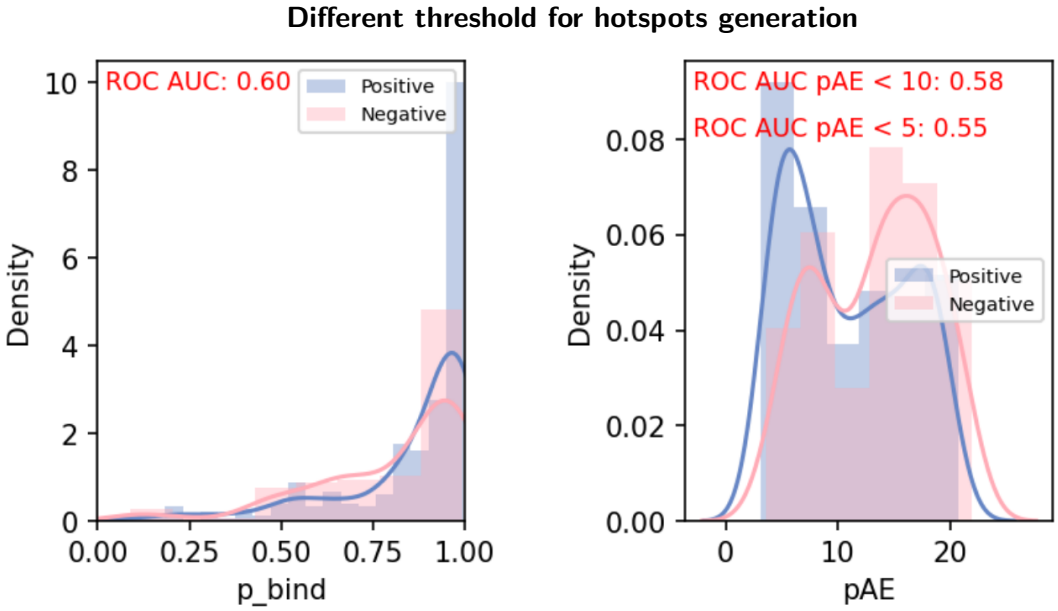
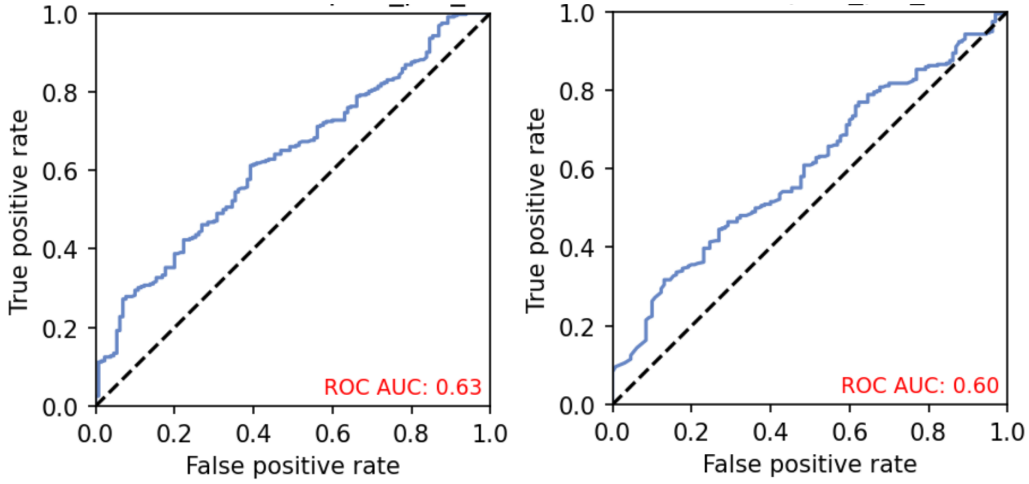


Figure 3.14: Distograms for two metrics: *pBind* and *pAE*



(a) The ROC curve for the *pAE* metric.

(b) The ROC curve for the *pBind* metric.

Figure 3.15: RFantibody was evaluated on two metrics; *pBind* and *pAE*. Changing the threshold for generating the hotspot during preprocessing of the training set led to a clear shift in the distribution of metrics between the binding and non-binding set, with AUCs for *pBind* and *pAE* being 0.60 and 0.63 respectively.

In comparison to the baseline model, a very similar trend can observed. However, this model which has a less strict margin for when a residue is considered a hotspot shows a slightly worse performance as the density plot in Figure 3.14 for the *pAE* suggests that the model has more false positives and false negatives. This is also reflected in the ROC curve in **??** where the AUC

is slightly lower than for the baseline model also showing that this approach did not yield a successful outcome.

3.2.5 Model with CDR Duplicate Removal

Here, the leading hypothesis was that maybe an additional bias was contained within the data set stemming from the sequence clustering which is based on sequence similarity. However, for antibodies this metric can be heavily skewed considering their high sequence similarity. Therefore the duplicates within the data set were removed using the CDR3 sequences alone, thus making the data set more strict and giving the model an easier task in distinguishing different binding events. The results of the training and benchmarking can be observed in Figure 3.16 and Figure 3.17.

Benchmarks

It was expected that this model would show a slightly better performance than the baseline model. However, most metrics displayed in Figure 3.16 and Figure 3.17 show no improvement in comparison to the baseline model and the model without fine-tuning on sequence-only data. For that reason, the hypothesis of another bias in the data set was not upheld.

3.2.6 Model without any Hotspots

To investigate the hypothesis, if the model is in fact utilizing information provided by the hotspots, the following approach was investigated. In order to compare to the other trials, the same sequence-based clustering was used. This model was trained on the same data set as the model in subsubsection 3.2.3, but without any hotspots provided. The results of the training and benchmarking can be observed in Figure 3.18 and Figure 3.19.

Benchmarks

When investigating the results of the benchmarking in Figure 3.18 and Figure 3.19 it becomes apparent that the model without any hotspots performs slightly worse than the baseline model with no fine-tuning. This is especially evident in the ROC curve in Figure 3.19 where the AUC for the *pBind* metric is significantly lower than for the other models. This suggests that the hotspots are indeed a necessary part of the model and removing them results in a worse performance - even though these hotspots were derived from the SSM data and not from the structure itself. Furthermore, *pAE* does not seem to worsen in comparison suggesting that the model is able to reason over the internal representation of the structure better than with the hotspots. For future investigations, the hotspot generation should be improved and varied to enhance the model's performance.

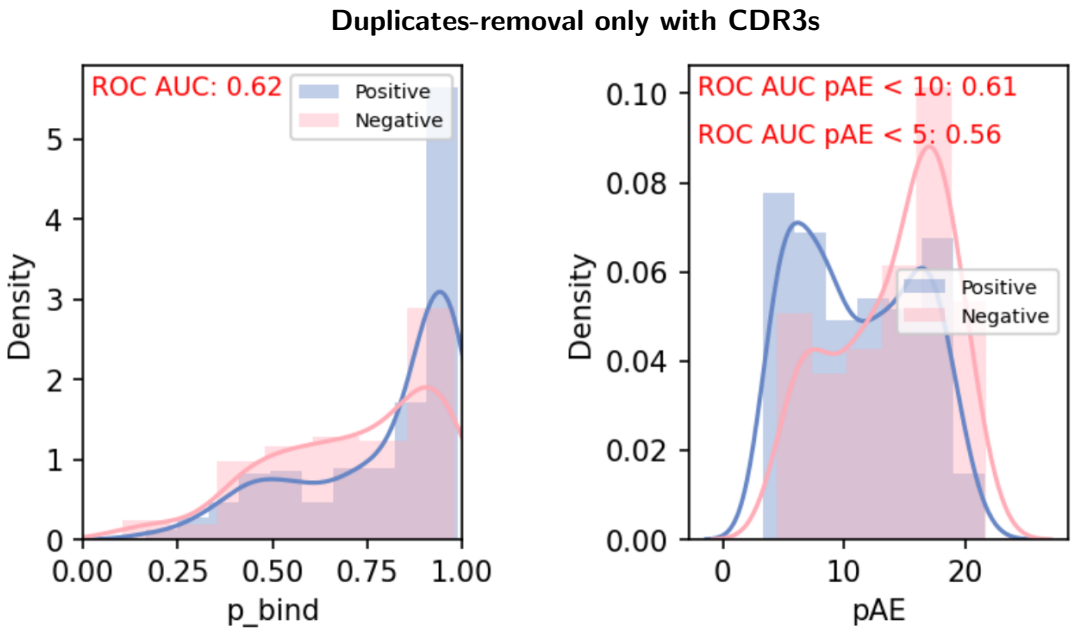
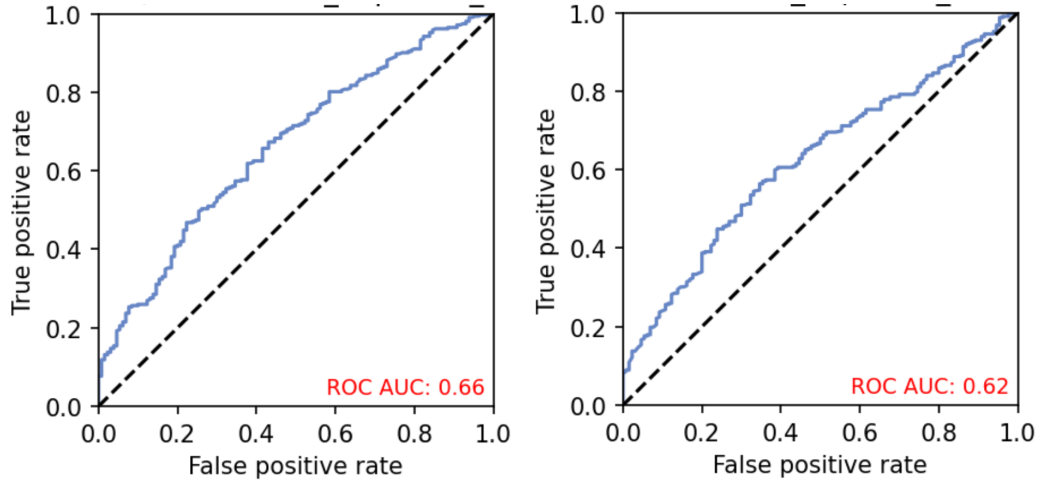


Figure 3.16: Distograms for two metrics: *pBind* and *pAE*



(a) The ROC curve for the *pAE* metric.

(b) The ROC curve for the *pBind* metric.

Figure 3.17: RFantibody was evaluated on two metrics; *pBind* and *pAE*. Removing duplicates using only the CDR3s before training led to a not very distinct difference in the distribution of metrics between the binding and non-binding set, with AUCs for *pBind* and *pAE* being 0.62 and 0.66 respectively.

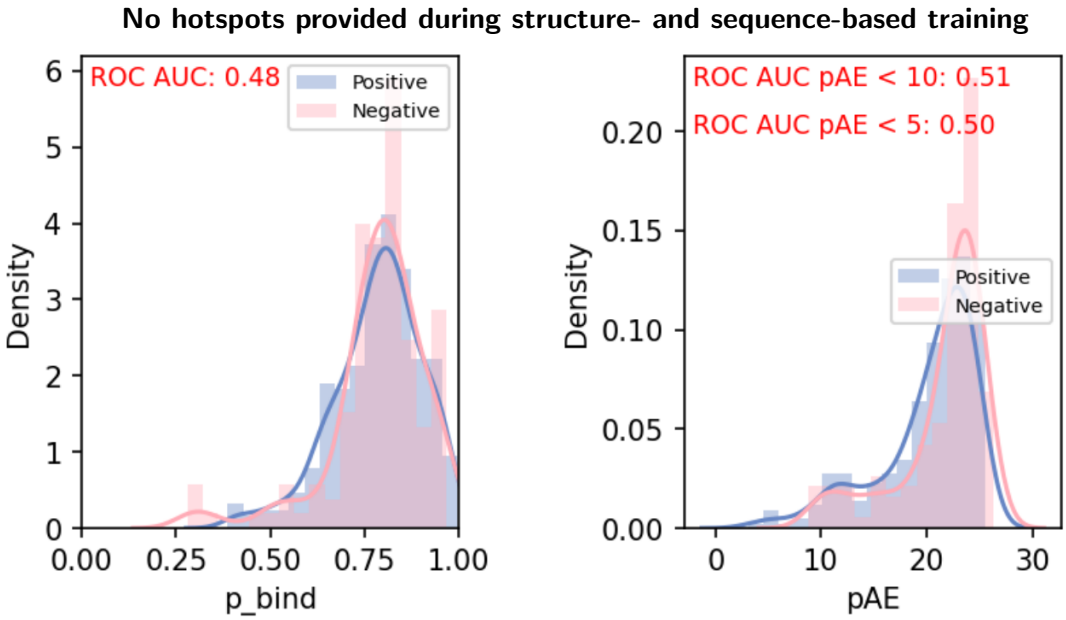
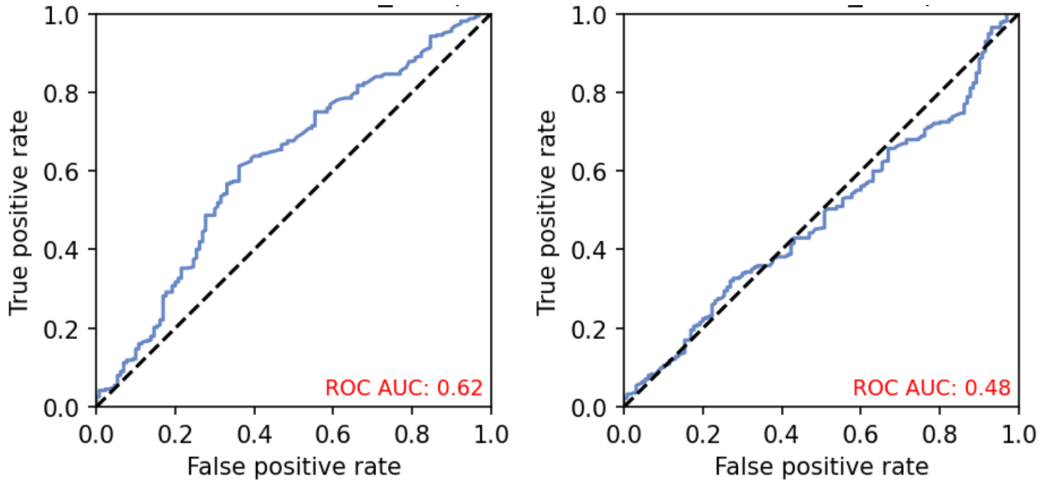


Figure 3.18: Distograms for two metrics: *pBind* and *pAE*



(a) The ROC curve for the *pAE* metric.

(b) The ROC curve for the *pBind* metric.

Figure 3.19: RFantibody was evaluated on two metrics; *pBind* and *pAE*. Not providing **any hotspots** at all during the training led to a distinct worsening in the distribution of metrics between the binding and non-binding set, with AUCs for *pBind* and *pAE* being 0.48 and 0.62 respectively.

4 Conclusion

We set out to improve antibody structure prediction given a sequence-only data set. In order to achieve that the model should first learn binding prediction. After identifying and correcting for a then discovered bias in the data set, a different clustering approach was employed and together with other experiments, the effects on performance were measured. However, changing the threshold for hotspot generation, or performing a more rigorous duplicate sequence removal for the preprocessing did not lead to improvements but degraded the model's performance on all tasks. This led us to test, if the model was able to correctly learn the information about the dock provided through the hotspots which we tested by omitting them. This showed that the model does indeed use the information contained within the hotspots since that specific model showed worse performance altogether. Sadly, we did not manage to show progress on the given task, despite attempting a less strict hotspot generation threshold, clustering in sequence space, being stricter with removing duplicates in the data set, and measuring the effect of omitting hotspots.

The general objective of this thesis was to investigate the feasibility of training RFantibody on a sequence-only data set. Much of the data presented here pertains to whether the model was learning to improve binding prediction I was training it on, which it did not, but the primary objective of whether training on sequence-only data can improve structure prediction remains to be seen. However, the work on this thesis has provided valuable insights into the challenges of generating a data set for training RFantibody on a sequence-only data set. The following section will highlight the key findings of this thesis and provide an outlook on the future of the project.

Improving the data set is indeed a critical step toward enhancing the model's performance. One of the more challenging issues was the bias introduced by the erroneous preprocessing and the imbalance of positive and negative binding events per residue position. Trying another sequence clustering approach, such as the specialized antibody similarity search method called Known Antibody Search (KA-Search) [53], could potentially provide a better clustering approach for the sequence based clustering than *MMseqs2* in order to obtain better data for training and validation. This can potentially mitigate biases and yield a more representative data set. However, the clustering performed by [38] better reflects the requirements for reasoning over structure potentially providing a better clustering approach overall.
Addressing the challenges related to hotspot derivation from the SSM data is important for accurate binding interface predictions. That information proved to be crucial for the correct prediction of the binding interface the training of RFantibody and correctly deriving the hotspots from the SSM data still remains challenging and different methods can be explored further.
As already explained in section 1.2 AF2 showed great improvement when trained with a self-distillation data set. Thus, a self-distillation data set for RFantibody, similar to the approach from AF2 or RF2 could enhance the model's robustness and generalization. By training RFan-

tibody on a diverse data set of antibodies, it may be possible to improve its performance and predictive capabilities significantly.
Moreover, incorporating another sequence-only data set of antibodies targeting a different antigen could be considered future steps. Varying the number of available structures and the proportion of different sequence-only antibody samples in this data set may prevent overfitting and offer more flexibility.
Lastly, integrating experimentally gathered data, such as from a large-scale high-throughput screening experiment focused on a diverse set of different targets, could provide the model with a diverse set of antibody interfaces to analyze and reason over. This could potentially enhance the model's predictive capabilities and generalization.

Bibliography

[1] Christian B. Anfinsen. "Principles that Govern the Folding of Protein Chains". en. In: *Science* 181.4096 (July 1973), pp. 223–230. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.181.4096.223. URL: https://www.science.org/doi/10.1126/science.181.4096.223 (visited on 11/13/2023).

[2] Carol A. Rohl et al. "Protein Structure Prediction Using Rosetta". en. In: *Methods in Enzymology*. Vol. 383. Elsevier, 2004, pp. 66–93. ISBN: 978-0-12-182788-5. DOI: 10.1016/S0076-6879(04)83004-0. URL: https://linkinghub.elsevier.com/retrieve/pii/S0076687904830040 (visited on 11/13/2023).

[3] Julia Koehler Leman et al. "Macromolecular modeling and design in Rosetta: recent methods and frameworks". In: *Nature methods* 17.7 (July 2020), pp. 665–680. ISSN: 1548-7091. DOI: 10.1038/s41592-020-0848-2. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7603796/ (visited on 10/18/2023).

[4] Andrew W. Senior et al. "Protein structure prediction using multiple deep neural networks in the 13th Critical Assessment of Protein Structure Prediction (CASP13)". en. In: *Proteins: Structure, Function, and Bioinformatics* 87.12 (2019). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10 pp. 1141–1148. ISSN: 1097-0134. DOI: 10.1002/prot.25834. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/prot.25834 (visited on 11/15/2023).

[5] Andriy Kryshtafovych et al. "Critical assessment of methods of protein structure prediction (CASP)—Round XIII". en. In: *Proteins: Structure, Function, and Bioinformatics* 87.12 (2019). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/prot.25823, pp. 1011–1020. ISSN: 1097-0134. DOI: 10.1002/prot.25823. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/prot.25823 (visited on 11/15/2023).

[6] Andrew W. Senior et al. "Improved protein structure prediction using potentials from deep learning". en. In: *Nature* 577.7792 (Jan. 2020), pp. 706–710. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/s41586-019-1923-7. URL: https://www.nature.com/articles/s41586-019-1923-7 (visited on 11/13/2023).

[7] Jianyi Yang et al. "Improved protein structure prediction using predicted interresidue orientations". en. In: *Proceedings of the National Academy of Sciences* 117.3 (Jan. 2020), pp. 1496–1503. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1914677117. URL: http://www.pnas.org/lookup/doi/10.1073/pnas.1914677117 (visited on 05/24/2021).

[8] Zeming Lin et al. "Evolutionary-scale prediction of atomic-level protein structure with a language model". en. In: *Science* 379.6637 (Mar. 2023), pp. 1123–1130. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.ade2574. URL: https://www.science.org/doi/10.1126/science.ade2574 (visited on 11/13/2023).

[9] John Jumper et al. "Highly accurate protein structure prediction with AlphaFold". en. In: *Nature* 596.7873 (Aug. 2021). Number: 7873 Publisher: Nature Publishing Group, pp. 583–589. ISSN: 1476-4687. DOI: 10.1038/s41586-021-03819-2. URL: https://www.nature.com/articles/s41586-021-03819-2 (visited on 10/23/2023).

[10] Fabian B. Fuchs et al. *SE(3)-Transformers: 3D Roto-Translation Equivariant Attention Networks*. arXiv:2006.10503 [cs, stat]. Nov. 2020. DOI: 10.48550/arXiv.2006.10503. URL: http://arxiv.org/abs/2006.10503 (visited on 11/16/2023).

[11] Karen Manalastas-Cantos et al. *Modeling flexible protein structure with AlphaFold2 and cross-linking mass spectrometry*. en. preprint. Biochemistry, Sept. 2023. DOI: 10.1101/2023.09.11.557128. URL: http://biorxiv.org/lookup/doi/10.1101/2023.09.11.557128 (visited on 11/16/2023).

[12] Georgios A. Pavlopoulos et al. "Unraveling the functional dark matter through global metagenomics". en. In: *Nature* 622.7983 (Oct. 2023). Number: 7983 Publisher: Nature Publishing Group, pp. 594–602. ISSN: 1476-4687. DOI: 10.1038/s41586-023-06583-7. URL: https://www.nature.com/articles/s41586-023-06583-7 (visited on 11/16/2023).

[13] *alphafold/docs/technical_note_v2.3.0.md at main · google-deepmind/alphafold · GitHub*. URL: https://github.com/google-deepmind/alphafold/blob/main/docs/technical_note_v2.3.0.md (visited on 11/16/2023).

[14] Richard Evans et al. *Protein complex prediction with AlphaFold-Multimer*. en. Pages: 2021.10.04.463034 Section: New Results. Mar. 2022. DOI: 10.1101/2021.10.04.463034. URL: https://www.biorxiv.org/content/10.1101/2021.10.04.463034v2 (visited on 11/16/2023).

[15] Mihaly Varadi et al. "AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models". en. In: *Nucleic Acids Research* 50.D1 (Jan. 2022), pp. D439–D444. ISSN: 0305-1048, 1362-4962. DOI: 10.1093/nar/gkab1061. URL: https://academic.oup.com/nar/article/50/D1/D439/6430488 (visited on 11/13/2023).

[16] Rohith Krishna et al. *Generalized Biomolecular Modeling and Design with RoseTTAFold All-Atom*. en. Pages: 2023.10.09.561603 Section: New Results. Oct. 2023. DOI: 10.1101/2023.10.09.561603. URL: https://www.biorxiv.org/content/10.1101/2023.10.09.561603v1 (visited on 11/13/2023).

[17] *A glimpse of the next generation of AlphaFold*. en. Oct. 2023. URL: https://deepmind.google/discover/blog/a-glimpse-of-the-next-generation-of-alphafold/ (visited on 11/14/2023).

[18] Minkyung Baek et al. *Efficient and accurate prediction of protein structure using RoseTTAFold2*. en. Pages: 2023.05.24.542179 Section: New Results. May 2023. DOI: 10.1101/2023.05.24.542179. URL: https://www.biorxiv.org/content/10.1101/2023.05.24.542179v1 (visited on 10/17/2023).

[19] Jeffrey A. Ruffolo et al. "Fast, accurate antibody structure prediction from deep learning on massive set of natural antibodies". en. In: *Nature Communications* 14.1 (Apr. 2023). Number: 1 Publisher: Nature Publishing Group, p. 2389. ISSN: 2041-1723. DOI: 10.1038/s41467-023-38063-x. URL: https://www.nature.com/articles/s41467-023-38063-x (visited on 11/16/2023).

[20] Jae Hyeon Lee et al. *EquiFold: Protein Structure Prediction with a Novel Coarse-Grained Structure Representation*. en. preprint. Bioinformatics, Oct. 2022. DOI: 10.1101/2022.10.07.511322. URL: http://biorxiv.org/lookup/doi/10.1101/2022.10.07.511322 (visited on 11/16/2023).

[21] Brennan Abanades et al. "ImmuneBuilder: Deep-Learning models for predicting the structures of immune proteins". en. In: *Communications Biology* 6.1 (May 2023). Number: 1 Publisher: Nature Publishing Group, pp. 1–8. ISSN: 2399-3642. DOI: 10.1038/s42003-023-04927-7. URL: https://www.nature.com/articles/s42003-023-04927-7 (visited on 11/16/2023).

[22] James Dunbar et al. "SAbDab: the structural antibody database". In: *Nucleic Acids Research* 42.D1 (Jan. 2014), pp. D1140–D1146. ISSN: 0305-1048. DOI: 10.1093/nar/gkt1043. URL: https://doi.org/10.1093/nar/gkt1043 (visited on 10/19/2023).

[23] Constantin Schneider, Matthew I J Raybould, and Charlotte M Deane. "SAbDab in the age of biotherapeutics: updates including SAbDab-nano, the nanobody structure tracker". In: *Nucleic Acids Research* 50.D1 (Jan. 2022), pp. D1368–D1372. ISSN: 0305-1048. DOI: 10.1093/nar/gkab1050. URL: https://doi.org/10.1093/nar/gkab1050 (visited on 10/19/2023).

[24] Philip Bradley. "Structure-based prediction of T cell receptor:peptide-MHC interactions". In: *eLife* 12 (), e82813. ISSN: 2050-084X. DOI: 10.7554/eLife.82813. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9859041/ (visited on 10/20/2023).

[25] Ruei-Min Lu et al. "Development of therapeutic antibodies for the treatment of diseases". en. In: *Journal of Biomedical Science* 27.1 (Dec. 2020). Number: 1 Publisher: BioMed Central, pp. 1–30. ISSN: 1423-0127. DOI: 10.1186/s12929-019-0592-z. URL: https://jbiomedsci.biomedcentral.com/articles/10.1186/s12929-019-0592-z (visited on 10/18/2023).

[26] Xiaochen Lyu et al. "The global landscape of approved antibody therapies". In: *Antibody Therapeutics* 5.4 (Sept. 2022), pp. 233–257. ISSN: 2516-4236. DOI: 10.1093/abt/tbac021. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9535261/ (visited on 10/18/2023).

[27] *Monoclonal Antibodies Market Size & Share Report, 2030*. en. URL: https://www.grandviewresearch.com/industry-analysis/monoclonal-antibodies-market (visited on 10/18/2023).

[28] Shijie Jin et al. "Emerging new therapeutic antibody derivatives for cancer treatment". en. In: *Signal Transduction and Targeted Therapy* 7.1 (Feb. 2022). Number: 1 Publisher: Nature Publishing Group, pp. 1–28. ISSN: 2059-3635. DOI: 10.1038/s41392-021-00868-x. URL: https://www.nature.com/articles/s41392-021-00868-x (visited on 10/19/2023).

[29] María Sofía Castelli, Paul McGonigle, and Pamela J. Hornby. "The pharmacology and therapeutic applications of monoclonal antibodies". en. In: *Pharmacology Research & Perspectives* 7.6 (2019). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/prp2.535, e00535. ISSN: 2052-1707. DOI: 10.1002/prp2.535. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/prp2.535 (visited on 10/18/2023).

[30] Meric Ovacik and Kedan Lin. "Tutorial on Monoclonal Antibody Pharmacokinetics and Its Considerations in Early Development". In: *Clinical and Translational Science* 11.6 (Nov. 2018), pp. 540–552. ISSN: 1752-8054. DOI: 10.1111/cts.12567. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6226118/ (visited on 10/19/2023).

[31] Bradley Brobst and Judith Borger. "Benefits and Risks of Administering Monoclonal Antibody Therapy for Coronavirus (COVID-19)". eng. In: *StatPearls*. Treasure Island (FL): StatPearls Publishing, 2023. URL: http://www.ncbi.nlm.nih.gov/books/NBK574507/ (visited on 10/18/2023).

[32] Joseph L. Watson et al. "De novo design of protein structure and function with RFdiffusion". en. In: *Nature* 620.7976 (Aug. 2023), pp. 1089–1100. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/s41586-023-06415-8. URL: https://www.nature.com/articles/s41586-023-06415-8 (visited on 10/17/2023).

[33] Nathaniel R. Bennett et al. "Improving de novo protein binder design with deep learning". en. In: *Nature Communications* 14.1 (May 2023). Number: 1 Publisher: Nature Publishing Group, p. 2625. ISSN: 2041-1723. DOI: 10.1038/s41467-023-38328-5. URL: https://www.nature.com/articles/s41467-023-38328-5 (visited on 10/18/2023).

[34] *Visualizing an invisible virus – The Pipettepen*. en. Mar. 2021. URL: https://www.thepipettepen.com/visualizing-an-invisible-virus/ (visited on 10/25/2023).

[35] 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". en. In: *Protein Science* 31.1 (Jan. 2022), pp. 141–146. ISSN: 0961-8368, 1469-896X. DOI: 10.1002/pro.4205. URL: https://onlinelibrary.wiley.com/doi/10.1002/pro.4205 (visited on 10/23/2023).

[36] Milot Mirdita et al. "Uniclust databases of clustered and deeply annotated protein sequences and alignments". en. In: *Nucleic Acids Research* 45.D1 (Jan. 2017), pp. D170–D176. ISSN: 0305-1048, 1362-4962. DOI: 10.1093/nar/gkw1081. URL: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkw1081 (visited on 10/23/2023).

[37] Martin Steinegger et al. "HH-suite3 for fast remote homology detection and deep protein annotation". en. In: *BMC Bioinformatics* 20.1 (Dec. 2019). Number: 1 Publisher: BioMed Central, pp. 1–15. ISSN: 1471-2105. DOI: 10.1186/s12859-019-3019-7. URL: https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-019-3019-7 (visited on 10/21/2023).

[38] Yunlong Cao et al. "Imprinted SARS-CoV-2 humoral immunity induces convergent Omicron RBD evolution". en. In: *Nature* 614.7948 (Feb. 2023). Number: 7948 Publisher: Nature Publishing Group, pp. 521–529. ISSN: 1476-4687. DOI: 10.1038/s41586-022-05644-7. URL: https://www.nature.com/articles/s41586-022-05644-7 (visited on 10/21/2023).

[39] *Welcome to Python.org*. en. Nov. 2023. URL: https://www.python.org/ (visited on 11/10/2023).

[40] The pandas development team. *pandas-dev/pandas: Pandas*. Nov. 2023. DOI: 10.5281/ZENODO.3509134. URL: https://zenodo.org/doi/10.5281/zenodo.3509134 (visited on 11/11/2023).

[41] Charles R. Harris et al. "Array programming with NumPy". en. In: *Nature* 585.7825 (Sept. 2020), pp. 357–362. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/s41586-020-2649-2. URL: https://www.nature.com/articles/s41586-020-2649-2 (visited on 11/10/2023).

[42] John D. Hunter. "Matplotlib: A 2D Graphics Environment". In: *Computing in Science & Engineering* 9.3 (2007), pp. 90–95. ISSN: 1521-9615. DOI: 10.1109/MCSE.2007.55. URL: http://ieeexplore.ieee.org/document/4160265/ (visited on 11/10/2023).

[43] Michael Waskom. "seaborn: statistical data visualization". In: *Journal of Open Source Software* 6.60 (Apr. 2021), p. 3021. ISSN: 2475-9066. DOI: 10.21105/joss.03021. URL: https://joss.theoj.org/papers/10.21105/joss.03021 (visited on 11/12/2023).

[44] Adam Paszke et al. "PyTorch: An Imperative Style, High-Performance Deep Learning Library". In: (2019). Publisher: arXiv Version Number: 1. DOI: 10.48550/ARXIV.1912.01703. URL: https://arxiv.org/abs/1912.01703 (visited on 10/28/2023).

[45] Ilya Loshchilov and Frank Hutter. "Decoupled Weight Decay Regularization". In: (2017). Publisher: arXiv Version Number: 3. DOI: 10.48550/ARXIV.1711.05101. URL: https://arxiv.org/abs/1711.05101 (visited on 10/28/2023).

[46] *Weights & Biases*. en. URL: https://wandb.ai/site (visited on 10/30/2023).

[47] Minkyung Baek et al. *Accurate prediction of nucleic acid and protein-nucleic acid complexes using RoseTTAFoldNA*. en. preprint. Bioinformatics, Sept. 2022. DOI: 10.1101/2022.09.09.507333. URL: http://biorxiv.org/lookup/doi/10.1101/2022.09.09.507333 (visited on 10/23/2023).

[48] Yunlong Cao et al. "Imprinted SARS-CoV-2 humoral immunity induces convergent Omicron RBD evolution". en. In: *Nature* (Dec. 2022). ISSN: 0028-0836, 1476-4687. DOI: 10.1038/s41586-022-05644-7. URL: https://www.nature.com/articles/s41586-022-05644-7 (visited on 10/23/2023).

[49] Laurens van der Maaten and Geoffrey Hinton. "Visualizing Data using t-SNE". In: *Journal of Machine Learning Research* 9.86 (2008), pp. 2579–2605. URL: http://jmlr.org/papers/v9/vandermaaten08a.html.

[50] Martin Steinegger and Johannes Söding. "MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets". en. In: *Nature Biotechnology* 35.11 (Nov. 2017), pp. 1026–1028. ISSN: 1087-0156, 1546-1696. DOI: 10.1038/nbt.3988. URL: https://www.nature.com/articles/nbt.3988 (visited on 10/27/2023).

[51] Kristian W. Kaufmann et al. "Practically Useful: What the Rosetta Protein Modeling Suite Can Do for You". In: *Biochemistry* 49.14 (Apr. 2010). Publisher: American Chemical Society, pp. 2987–2998. ISSN: 0006-2960. DOI: 10.1021/bi902153g. URL: https://doi.org/10.1021/bi902153g (visited on 10/18/2023).

[52] Jun Lan et al. "Structure of the SARS-CoV-2 spike receptor-binding domain bound to the ACE2 receptor". en. In: *Nature* 581.7807 (May 2020), pp. 215–220. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/s41586-020-2180-5. URL: https://www.nature.com/articles/s41586-020-2180-5 (visited on 10/28/2023).

[53] Tobias H. Olsen et al. "KA-Search, a method for rapid and exhaustive sequence identity search of known antibodies". en. In: *Scientific Reports* 13.1 (July 2023), p. 11612. ISSN: 2045-2322. DOI: 10.1038/s41598-023-38108-7. URL: https://www.nature.com/articles/s41598-023-38108-7 (visited on 10/29/2023).