Enhancing Protein Sequence Design through Augmented Machine Learning of Hydrogen Bonding Networks

Citation

Permanent link
https://nrs.harvard.edu/URN-3:HUL.INSTREPOS:37379285

Terms of Use
This article was downloaded from Harvard University’s DASH repository, and is made available under the terms and conditions applicable to Other Posted Material, as set forth at http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA

Share Your Story
The Harvard community has made this article openly available. Please share how this access benefits you. Submit a story.

Accessibility
Enhancing Protein Sequence Design through Augmented Machine Learning of Hydrogen Bonding Networks

Kevin Tan

A thesis presented to
Computer Science
in partial fulfilment of the honors requirements for the degree of Bachelor of Arts
Abstract

Creating proteins that bind tightly and specifically to ligands is a significant challenge for protein sequence prediction. Current machine learning models struggle to design hydrogen bonding networks in proteins which are crucial for structure stability and ligand affinity. In this thesis, we explore how data on these higher-order interactions can better inform binding site design. We discuss how buried polar residues form interactions with their environment similar to ligands bound to proteins. We then present a strategy to augment training data with diverse, robust examples of hydrogen bonding networks satisfying these residues. This data is used to train a graph neural network that selects residues to explicitly model as standalone ligands. The model analysis demonstrates that predicted binding site sequences establish more realistic interactions with ligands, even for held-out classes of proteins. This suggests that biasing learning toward hydrogen bonding networks using buried residues can improve the performance of de novo sequence design.
Acknowledgements

I would like to thank my advisor, Dr. Nicholas Polizzi, for seeing my potential and providing continuous support throughout this thesis. Without his generosity, dedication, and insight, this research would not have been possible. I also thank Benjamin Fry for the time and energy he has invested in guiding me through the LASErMPNN project and for making his software accessible. I thank the members of the Polizzi lab for always making me feel welcome in the space. Thank you to Dr. Chris Sander, Dr. Hanspeter Pfister, and Dr. Wanhua Li for volunteering their time and serving as a thesis reader. I would like to thank all the friends I have met during my extended journey at Harvard. Thank you for pushing me to grow in so many ways and for sharing your warmth, laughter, and memories. Finally, I thank my parents and my sister for being the people I can always return home to. I hope to make you proud.
# Contents

<table>
<thead>
<tr>
<th>List of Figures</th>
<th>viii</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Introduction</td>
<td>1</td>
</tr>
<tr>
<td>1.1 Contributions</td>
<td>3</td>
</tr>
<tr>
<td>1.2 Related Work</td>
<td>4</td>
</tr>
<tr>
<td>2 Background</td>
<td>6</td>
</tr>
<tr>
<td>2.1 Protein Structure Determinants</td>
<td>7</td>
</tr>
<tr>
<td>2.2 Ligand Binding Proteins</td>
<td>8</td>
</tr>
<tr>
<td>2.3 Databases and Informatics</td>
<td>9</td>
</tr>
<tr>
<td>3 Strategy</td>
<td>11</td>
</tr>
<tr>
<td>3.1 Designing Hydrogen Bonding Networks</td>
<td>11</td>
</tr>
<tr>
<td>3.2 Modeling Ligand Interactions</td>
<td>12</td>
</tr>
<tr>
<td>3.3 Augmenting Training Data</td>
<td>14</td>
</tr>
<tr>
<td>4 Pseudoligand Candidate Database</td>
<td>17</td>
</tr>
<tr>
<td>4.1 Defining Metrics for Good Pseudoligands</td>
<td>17</td>
</tr>
<tr>
<td>4.1.1 Hydrogen Bonding Contacts</td>
<td>19</td>
</tr>
<tr>
<td>4.1.2 Burial</td>
<td>20</td>
</tr>
<tr>
<td>4.1.3 Conservation and Diversity</td>
<td>21</td>
</tr>
<tr>
<td>4.2 PDB Candidates Approach</td>
<td>23</td>
</tr>
</tbody>
</table>
List of Figures

3.1 Overview of Pseudoligand Augmentation Strategy ......................... 15

4.1 Visualization of Metrics for a Good Pseudoligand ......................... 18

4.2 Distribution of Pseudoligand Metrics ................................... 29

4.3 Breakdown of Pseudoligand Candidate Counts ............................ 31

5.1 LASErMPNN Performance with Pseudoligand Candidates ............... 37
Chapter 1

Introduction

Protein design is an active area of research within the field of computational biology, with the goal of recovering or generating amino acid sequences that fold to protein structures achieving a specific function. The ability to more accurately predict sequence from structure has led to advancements within fields like bioengineering and biomedicine because proteins can now be designed from scratch or created using naturally occurring structures as templates [1]. For instance, as antibiotic resistance limits the number of effective remedies available to treat diseases, the design of antibodies that specifically target pathogens provides an additional tool for antibiotic discovery and can lead to the creation of therapeutics that have fewer off-target side effects [2]. Industrial synthetic chemistry processes may benefit from enzymes that serve as novel biological catalysts to speed up chemical reaction rates [3]. Biosensors can be created using proteins that specifically bind to and detect molecules, opening up avenues for the creation of diagnostic and analytical tools [4].

The close tie between function and structure in proteins and the infinitely large sample space of possible sequences necessitates a rational approach to discovering novel, viable sequences [5, 6]. Machine learning techniques offer new ways to improve protein design by leveraging the vast amounts of data available on protein structure and relationships [7]. Popular software tools such as AlphaFold [8], RosettaFold [9], and ESMFold [10] leverage neural networks trained on experimentally determined protein structures to deliver groundbreaking accuracy for protein structure predictions, and applications to the inverse problem of protein sequence design have also obtained increases in
accuracy [11, 12].

Despite these advancements in part due to the application of machine learning models, it is still challenging to design sequences that fold to structures binding small molecules [13]. More generalized approaches to de novo protein design have yielded models that can confidently predict large-scale dynamics that drive protein structure formation, such as the packing of hydrophobic residues in the center of a protein or the use of common structural motifs to stabilize the protein [12]. However, these considerations may overshadow the fine-grain interactions that define protein-ligand binding, particularly when polar interactions are needed at the binding site [14]. The challenge is two-fold. For one, these structures must be stable both in the presence and the absence of the ligand [15], meaning that residues in the binding pocket have the dual task of providing structure and affinity. Because many machine learning models are trained on proteins that exclude ligands or only include monomeric structures, the success rate for such computational predictions is low, resulting in the need to conduct additional high-throughput screening [16]. In addition, hydrogen bonding networks are a crucial interaction for ligand specificity and affinity [17], but the inclusion of polar residues within binding sites must result in particular energetically favorable interactions to overcome the hydrophobic effect [18]. Thus, machine learning models must learn to include and satisfy these polar interactions during protein sequence design.

While machine learning offer a path forward for incorporating these biological constraints in prediction, classic problems associated with machine learning emerge even in the state-of-the-art models. The training of these models on comprehensive datasets such as the Protein Data Bank leads to questions of data leakage between training and evaluation. These models typically perform well in scenarios where the predicted protein shares similarities with families of proteins that have experimentally validated structures but struggle with rarer structural motifs or artificial structures not found in nature [19, 20]. As such, these models are less reliable for applications such as the design of ligand-binding proteins because small molecules are often excluded from the training process. Adding new data is also expensive and time-consuming because of the difficulty of preparing proteins for analysis, and structures for protein-ligand complexes are particularly difficult to acquire [21]. Synthetic data is not easily accessible because of limitations in the current understanding of protein
folding [22].

Through this thesis, we seek to improve protein sequence design for ligand-binding proteins by augmenting the data available for the training of machine learning models. The exclusion of non-protein molecules in the training of these models results in a gap in the data for the hydrogen bonding interactions that are crucial for facilitating both structural resilience and ligand binding. Because empirically adding structures for new protein complexes is difficult, we are motivated to develop an alternative method to do more with existing available data. Specifically, we are motivated by the use of amino acid sidechains as a proxy for small molecules, using interactions found in protein structures to predict how proteins can build hydrogen bonding interactions around ligands with similar chemical groups. Focusing a machine learning model’s attention on the building blocks of folding and binding may introduce crucial domain knowledge that improves accuracy for small-molecule binding protein sequence predictions.

1.1 Contributions

This thesis investigates how the creation of a dataset centered around hydrogen bonding interactions in proteins can inform machine learning models about how to build hydrogen bonding networks around small-molecule ligands. We discuss factors that are important for protein folding and how designing for ligand-binding proteins is especially challenging. Limitations to the amount of high quality training data available for the task designing binding sites motivates an approach for data augmentation that uses existing experimental data on protein structures as well as computationally predicted structures, with a focus on interactions between a small-molecule and its residue environment.

The central contribution is an analysis across the deposited structures in the Protein Data Bank to find and rank examples of residues that are highly likely to participate in hydrogen bonding networks similar to those needed for protein-ligand binding. Residues are filtered and scored using computed information about the residue’s solvent-accessible surface area, conservation across the multiple sequence alignment, and hydrogen bonding interactions. Using multiple sequence
alignments, these residues are compared across structures that share homology, incorporating diverse networks that bind to a particular sidechain. Furthermore, the AlphaFold Protein Structure Database is leveraged to add protein structures that have not yet been crystallized, establishing a novel dataset for how amino acid residues establish higher-order binding interactions in proteins.

Finally, we compare and evaluate the performance of machine learning models using this database. This dataset is incorporated into a graph neural network called LASErMPNN (Ligand Aware Sequence EngineerRing MPNN) that samples sidechains as pseudoligands, augmenting the model’s knowledge about potential sequences that bind to that chemical group. We introduce preferential sampling of high-scoring pseudoligand candidates, and we show that biasing machine learning toward hydrogen bonding networks can improve binding site design without adversely affecting overall structural prediction accuracy. Through this inclusion, we hope to create a sequence prediction model that can accurately model the more nuanced interactions that guide protein folding and ligand binding, allowing for the rapid development of empirically viable, small-molecule protein binders.

Chapter 2 of this thesis provides background on protein structure determinants and introduces terminology specific to the problem setting. Chapter 3 outlines the strategy of sampling buried residues as pseudoligands and describes the considerations that are undertaken by this approach. Chapter 4 discusses the theoretical framework and implementation of a residue scoring system to find pseudoligands that engage with hydrogen bonding networks, describing the features measured and the contents of the compiled database. Chapter 5 demonstrates the use of this pseudoligand database in a graph neural network, details the architecture of the model and how residues are sampled, and discusses the experimental findings. Finally, chapter 6 concludes and mentions areas of future work.

1.2 Related Work

Deep-learning approaches to protein sequence prediction have been accomplished with the application of graph neural networks, an architecture that more naturally captures the bonds between
the residues in a sequence. ProteinMPNN [23] is one such graph neural network that improves upon physics-based approaches such as Rosetta [24], offering a sequence recovery of 52.4% on native protein backbones compared to 32.9% for Rosetta. This approach of recovering sequences from a backbone is useful to the field of protein design because it enables the design of sequences that fold to a particular structure and thus take on a particular function. The ProteinMPNN model is trained on sequence recovery of single-chain proteins from the PDB and thus achieves high accuracy benchmarks on native protein sequence recovery. However, for novel protein design tasks, it works best when rescuing previously failed designs from other protein prediction models and is limited when designing multimeric proteins, structures that have no natural equivalents, or explicit ligands. Nonetheless, ProteinMPNN shows promise with faster inference times compared to traditional physical sequence design models and higher rates of experimental design success [23].

The closest related work surrounding sequence design for ligand-binding proteins comes from LigandMPNN [25], an expansion to ProteinMPNN that explicitly models non-protein atoms and small molecules. The introduction of these non-protein elements to the deep-learning model allows for the prediction of residues that form interactions with ligands because these inputs can be represented as a separate ligand graph that is connected to the protein graph. Ligand inputs are passed through additional encoder layers to model protein-ligand interactions, and information is passed to residues that are the nearest neighbors to the ligand to influence residue identity and rotamer conformation in the sequence. During the training of LigandMPNN, the dataset of ligands is augmented by randomly sampling 2-4% of protein residues to use their sidechains as ligand examples in addition to other small molecules, nucleotides, and metals in the experimental data. This data augmentation allowed the addition of atoms from amino acids to ligand graphs to stabilize binding sites but did not significantly increase sequence recovery [25]. The authors suggest that rotamer predictions gain most information from protein binding site pre-organization rather than from added ligand context, which indicates that additional information around ligand neighborhoods can improve model performance.
Chapter 2

Background

The "protein folding problem" is a long-standing area of research that seeks to understand the relationship between a protein’s amino acid sequence to its structure [22]. Protein sequence design is a subset of this field that focuses on generating sequences that natively fold to a particular three-dimensional structure. The Anfinsen hypothesis states that the thermodynamic stability of a native structure stems from its sequence and the interaction with its environment [26], which motivates the development of structure-to-sequence models that learn the rules and relationships governing protein folding. This task is difficult since de novo protein design often involves designing around a target function, of which multiple structures can serve as a solution; furthermore, protein structures have many degrees of freedom that make it difficult to predict how sequences will fold in vivo. This phenomenon coined as Levinthal’s paradox suggests that proteins do not occupy all possible spatial configurations and are instead biased toward motifs due to physical interactions [27]. As such, a focus on these interactions critical for protein folding can help us understand how to better predict sequences that lead to viable protein structures [28]. Machine learning models, which can be attuned to the patterns underlying the data, can benefit from training data that includes representative and diverse interactions that lead to stable and functional protein structures.
2.1 Protein Structure Determinants

Protein sequences consist of amino acids linked together linearly, and each amino acid consists of the backbone component, which forms the main chain, and the sidechain component, which determines the amino acid’s unique chemical properties. The interactions between the chemical groups in amino acids drive protein folding by establishing thermodynamically favorable states [22]. It is thought that in the native state of a protein, the free energy of the structure is minimized through local interactions with amino acid residues adjacent in sequence and through global interactions that are the result of larger-scale forces [29].

The dominant driving force of protein folding is hydrophobicity, a physical property of nonpolar amino acids that repels them from water and polar amino acids and attracts them to other nonpolar amino acids. Because proteins are synthesized in and exposed to their environment, often containing water, nonpolar residues in the protein are driven to the core of the structures to minimize unfavorable interactions with water [30]. The packing of nonpolar amino acids maximizes the hydrophobic interactions between residues while making it unfavorable for water molecules to enter the protein structure and introduce polar interactions. The difference in free energy because of this positioning of nonpolar and polar regions in a protein contributes to the stability of a protein fold and is thus an extremely common interaction in a protein structure [29].

Hydrogen bonding is another critical interaction for protein stability, occurring between polar residues that have electronegative atoms such as nitrogen or oxygen. The partial charges from these residues result in hydrogen bonds with other polar residues or with water, and because unsatisfied hydrogen bond donors or acceptors can weaken binding between 1-4 kcal/mol [31], hydrogen bonding between polar amino acids can be incredibly stabilizing in environments where an amino acid residue is not exposed to a polar solvent. Hydrogen bonding can occur with an amino acid’s backbone or sidechain group, but while backbone hydrogen bonding interactions can stay unchanged during an amino acid mutation, a change in the sidechain group can lead to vastly different native structures, as in the case when a polar amino acid is changed to a nonpolar one [31]. Because of this, the identity of an amino acid residue and the placement of polar sidechains
are important for understanding how a sequence folds to its final structure [32].

Other interactions contributing to protein structure include Van der Waals forces and packing, both of which are related to the relative distance between amino acids. London dispersion forces, which are a type of Van der Waals force, result from induced dipoles in atoms which lead to small, temporary attractions between amino acids. Atoms that are packed tightly together can experience these forces, and while London dispersion forces are small individually, they contribute in aggregate to the stability of the protein [33]. Conversely, amino acids cannot be packed too closely since clashing can occur when atoms are forced to occupy the same position in space. An amino acid’s sidechain has a limited range through which it can rotate in three-dimensional space, and structures that fit residue sidechains together neatly, much like puzzle pieces, are energetically preferred. The positioning of these sidechains, called rotamers, induces conformational changes through clashing and influences the strength of directional interactions such as hydrogen bonds, which strengthen when the partial charges on an atom are aligned [34].

These considerations are paramount in the creation of a sequence for a given structure because they represent important physical interactions that drive protein folding. On a large scale, hydrophobic packing forces dominate in deciding where amino acids are placed, but hydrogen bonding, Van der Waals forces, and rotamer placement are fine-grain interactions that are just as influential in the specifics of protein structure determination and, as we will later see, in the function of a protein.

### 2.2 Ligand Binding Proteins

There are additional constraints on protein sequence design when the protein aims to bind to ligands, which are molecules not part of the main protein sequence that form a complex with the protein. Proteins will often have a binding pocket where specific amino acids create a favorable environment for the ligand to fit the protein by using polar and nonpolar interactions, hydrogen bonding, and shape [15]. The specificity of the protein to the ligand and the affinity for the protein-ligand interaction are determined by the placement of such amino acids. In the absence
of the ligand, the amino acids that drive ligand binding must still be satisfied, meaning that any potential unfavorable interactions must be stabilized by solvent or other amino acids [15]. For polar ligand binding sites, it is important that the amino acids in the site form interactions that are stronger than the hydrophobic forces that may work to unravel the protein.

Challenges may arise in the structural differences between bound and unbound states. Protein sequences designed using computational techniques often require extensive experimental screening to discover which structures are viable, because sequence predictions often fail to fold to a structure that is stable or has a high affinity for the ligand [35]. In addition, the binding process can have variations depending on how ligands interact with the binding site, and multiple conformational changes from the protein may be needed during the process of fitting a ligand to a site [35]. The difference in free energy between bound and unbound protein structures is often much smaller than the difference between folded and unfolded proteins because of the small number of interactions that facilitate the binding process. As a result, this makes it difficult to computationally model how proteins accomplish tight and specific binding, despite these interactions being so critical to real-world performance [35].

2.3 Databases and Informatics

The field of protein prediction is built off a vast amount of data, a culmination of decades of research and expedited by the evolution of high-resolution, high-throughput processing technologies. The Protein Data Bank (PDB) is a repository of over 100,000 unique experimentally determined protein structures that contain information about the atomic coordinates of both proteins and complexes that include DNA, RNA, or small molecules [36]. For each protein sequence, a multiple sequence alignment (MSA) can be generated, which aligns and ranks groups of protein sequences based on their assumed evolutionary relationship [37]. Protein sequences can change with the insertion, deletion, or mutation of an amino acid, and an MSA aims to align similar parts of the sequence together, predicting which amino acids or sections of the sequence may have undergone these changes. Columns of the MSA are thought to correspond with amino acids that share the same
lineage. As such, phylogeny can be tracked through an MSA, demonstrating how evolutionary changes manifest in a sequence, and revealing additional information about amino acid changes that occur in tandem and over time. Amino acids that are conserved across the MSA may indicate that an amino acid is stabilized by its neighboring residues. Alternatively, having various amino acids across a column in the MSA can represent significant changes across the protein sequences, with either point mutations changing the identity of the amino acid at that position or insertions and deletions shifting the sequence [37].

The relative ordering within a protein sequence and the three-dimensional structure is important to consider, especially when looking for interactions that govern stability or ligand specificity. Protein structures contain information about the coordinates of the atoms within a folded protein, and additional information can be extracted from these structures that are not present within just the sequence. For one, the placement of an amino acid relative to the surface or the core of a protein affects its exposure to polar and nonpolar interactions. The degree to which an amino acid is buried within a protein can be measured using its solvent-accessible surface area (SASA), which measures the surface area of the molecule exposed to the environment [38]. In addition, the intermolecular interactions available to an amino acid depend on which molecules are within a certain contact distance. Because of this, the neighborhood that the amino acid is located in, determined by the residues that fall within the contact radius of the amino acid, is instrumental in providing stabilizing interactions. For hydrogen bonding, that contact radius is typically 3.5 angstroms, while other heavy atom interactions can occur as far as five angstroms away from the residue [39].
Chapter 3

Strategy

3.1 Designing Hydrogen Bonding Networks

Hydrogen bonds are an important intermolecular interaction for protein structures, and the polar amino acid residues that facilitate these bonds can be used to shield hydrophobic residues in the core of the protein from polar solvents [30], contribute to internal structural specificity [31], or provide function through binding sites [15]. These interactions are important for designing sequences that target specific functions and for the performance and production of these proteins. De novo protein designs that rely solely on hydrophobic packing for structure tend to have lower expression, can precipitate during purification, and may fold to different structures than what is predicted [40]. Protein designs that satisfy all hydrogen bond donors and acceptors are more likely to be successful than protein designs that have less robust hydrogen bonding [41]. Thus, when training a machine learning model for protein sequence design, the model needs to incorporate hydrogen bonding into its considerations and predictions.

The value of hydrogen bonding contrasts with the difficulty in successfully designing these higher-order interactions. Hydrogen bonds are strongest when the hydrogen bond donor and the hydrogen bond acceptor are pointed directly at one another [42] and when the atoms have 2.5 to 3.2 angstroms between them [43]. Limitations to the rotamers that a residue can take on means that these polar residues are rarely satisfied by a single hydrogen bond and instead participate in larger
hydrogen bonding networks where multiple residues contribute partially to satisfy a polar residue [44]. These particularities result in the need for highly precise placement of polar amino acids to create favorable interactions, as the sidechains themselves must be oriented in a favorable rotamer and have the correct placement in the protein structure [32, 45]. In addition, for ligand binding sites where hydrogen bonds are key for protein-ligand specificity and affinity, amino acids must establish those networks in various conformations [46, 47], including when the protein is bound and unbound to the ligand.

Because the hydrophobic packing that stabilizes protein interactions on a coarse scale is easier to generalize, many machine learning models over-rely on hydrophobic interactions and lack hydrogen bonds that improve structural stability or lead to better ligand binding [35, 45]. The difficulty of designing hydrogen bonding networks from scratch and the penalization of unsatisfied polar residues within the loss of many protein design models are in opposition to each other and thus present a significant challenge for machine learning models. A key addressable component is the presence of hydrogen bonding networks across the training data, which is limited by the relative rarity of successful networks within proteins [48]. In addition, prediction models that are not focused on the task of ligand-binding protein sequence design often exclude ligands in their input data, further limiting the number of hydrogen bonding interactions available during training. Even ligand datasets are limited in size due to the difficulty of obtaining empirical structures of complexes [49]. This motivates the explicit modeling of protein-ligand interactions within the training of protein prediction models and a focus on existing successful hydrogen bonding interactions and networks within protein datasets.

## 3.2 Modeling Ligand Interactions

Advances within protein sequence design methodologies have led to improved performance for ligand-binding predictions. For instance, ProteinMPNN [23] is a deep learning protein sequence design method that achieves higher sequence recovery on native protein backbones compared to previous physics-based methodologies such as Rosetta [24]. Frameworks leveraging machine learning
can generate solutions with significantly fewer compute resources and with higher sequence recovery [50]. One limitation of ProteinMPNN is that it does not account for protein-nucleic acid or protein-small molecule design; however, this has been addressed by models such as LigandMPNN [25], which extends the ProteinMPNN architecture while including the ability to model non-amino acid molecules by explicitly modeling ligands. This is a step in the right direction for the field of protein sequence design, but as with many other machine learning models trained over PDB data, both struggle when the ligand or the protein is novel and not found in the PDB [25]. In other words, additional work must still be made to improve prediction performance for de novo ligand-binding protein designs. It is suggested that the limited high-quality training data available for protein-small molecule binding is a bottleneck for improved performance [48, 51].

The design of hydrogen bonding networks is important to confer specificity between the protein and the ligand [31], and work has been done to improve the fidelity of sequence prediction for the task of ligand binding. A van der Mer (vdM) is a unit of protein structure that relates the positioning of interacting chemical groups of an amino acid with the positioning of the atoms that residue's backbone atoms [52]. These vdMs are sampled from PDB structural data and are used to predict statistically likely positioning, allowing proteins to be designed around binding to small molecules rather than simply predicting protein structure. vdMS can identify binding-site residues that can form hydrogen bonds to small molecules, and the COMBS algorithm can be used to identify highly favorable interactions like solutions found in existing PDB structures [52]. This suggests that knowledge about hydrogen bonding networks, particularly those commonly found in protein structure motifs, can be used to assist in the task of protein design.

The establishment of databases of good ligand-protein interactions such as van der Mers inspires a need for additional data for the task of designing binding sites with structurally stable and ligand-specific amino acids. Considering the difficulty of gathering more empirical protein complex data, a method of informed synthetic data augmentation is needed to achieve better performance with more ligand-binding training data. Drawing from the approach of constructing vdMs, small-molecule ligands can be decomposed into more common chemical groups, some of which share similarities with chemical groups in amino acid sidechains. Additional solutions to the ligand binding problem
may be found in the construction of stabilizing interactions with buried polar residues, which share both desirable chemical properties, size, and interactions. Using these residues as examples of ligand-bound structures can be fruitful for discovering variable neighboring residue conformations that are energetically favorable for a chemical group ligand.

### 3.3 Augmenting Training Data

Sequence predictions for machine learning models seeking to design ligand-binding proteins can be improved by directing learning toward higher-order interactions that are difficult to design or generalize, the hydrogen bonding networks that are established at ligand-binding sites or within proteins. To do so, we construct a novel dataset that identifies residues within protein structures that participate in hydrogen bonding networks like those of ligand-protein complexes. As in the van der Mers approach where molecules are broken down into their component chemical groups, we use a residue’s sidechain and the neighboring residues it comes into contact within the protein structure as an example of a ligand interacting with the binding site of a protein. These selected residues can serve as "pseudoligands" during the model training process, introducing novel "ligand" binding sites from which key interactions and relations can be learned. By biasing learning toward establishing hydrogen bonding networks in sequences, binding site predictions can draw from the sequence space not available in existing ligand-protein datasets and thus better design for small-molecule ligands that are novel or not found in nature.

Not all residues within a protein can provide helpful examples for training data, and a few desirable properties of a pseudoligand are outlined in Figure 3.1 A. For a residue sidechain to engage in a hydrogen bonding network that is relevant to ligand binding, the residue must have many potential hydrogen bonding residues within contacting distance and be buried in the protein to minimize exposure to solvent or participate in a real ligand binding site. In addition to these properties, we are interested in seeing how such residues participate in diverse residue environments, mimicking the diversity in binding site conformations available to real ligands. As such, having a highly conserved residue with a neighborhood that has a high diversity across their multiple
Figure 3.1: **Overview of Pseudoligand Augmentation Strategy.** A) Residues that serve as good pseudoligand candidates share environments like ligand binding sites B) A pipeline for finding diverse pseudoligands is established using multiple sequence alignments. Pseudoligands can be used to augment training data for sequence prediction machine learning models like LASErMPNN.
sequence alignment indicates that the residue participates in various slightly different hydrogen bonding networks, allowing us to find alternative solutions in nature for satisfying a particular sidechain group.

Finding residues that serve as good candidates for pseudoligands can introduce diverse examples of "binding site" interactions within the PDB that do not require the search for empirical ligand-protein structures. Pseudoligands and the hydrogen bonding networks they participate in can also be found beyond the PDB through the introduction of computationally generated structure, which can provide helpful structural information for many unresolved homologs in multiple sequence alignments. This adds yet another source of data diversity, allowing a larger set of successful interaction networks with amino acid sidechains. As illustrated in Figure 3.1 B, the search for good pseudoligand candidates within the PDB can be extended using multiple sequence alignments, using them as a heuristic for candidates within unresolved sequences. These pseudoligands can then be used to augment existing training data for machine-learning models that leverage ligand-protein datasets.
Chapter 4

Pseudoligand Candidate Database

In the previous chapter, we related the challenge of designing hydrogen bonding networks with that of designing ligand-binding proteins, showing how improving the performance of a model on predicting polar residue placement can also lead to more stable, higher affinity binding sites. We also demonstrate how buried residues and their interactions with their neighborhood can provide a novel source of data for how proteins form favorable environments in complex with small-molecule chemical groups. This motivates the creation of a database of residues that may serve as high-quality pseudoligands during the training of a ligand-binding protein design model. This pseudoligand candidate database increases the number of ligand-protein training examples available during training and offers quantifiable metrics that can be used during model training.

4.1 Defining Metrics for Good Pseudoligands

To establish a database of pseudoligand candidates, we define features that measure how a residue participates in buried hydrogen bonding networks and that indicate the diversity of alternative networks available across related sequences. The search and scoring of residues across protein structures are extended to sequences that have not yet been experimentally determined, using multiple sequence alignments and computationally generated structures to bridge the gap and augment training data with novel pseudoligand examples.
Figure 4.1: Visualization of Metrics for a Good Pseudoligand. A) The highlighted residue forms many hydrogen bonds with neighboring residues, marked by the dashed lines. B) The highlighted residue is buried by neighboring residues, indicated by the lack of gaps in the Gaussian surface of the neighborhood. C) The highlighted residues are residues in sequences in the MSA that correspond to the original pseudoligand residue. The diversity of residues across the neighborhoods of the three sequences provides alternative structural networks that bind to the residue.
4.1.1 Hydrogen Bonding Contacts

The primary goal of our search is to discover unique examples of hydrogen bonding networks around a candidate residue (Figure 4.1 A). This diversity can arise from a change in the residues within contacting distance, making it important to track the number and types of potential hydrogen bonding pairings within a residue’s neighborhood. In particular, the hydrogen bonding contacts between the sidechain groups of residues in the protein inform how the arrangement of residue chemical groups forms a stable hydrogen bonding network, more so than contacts to the backbone of the residue which are constant across amino acids. We thus define a hydrogen bonding contact between residues when the nitrogen and oxygen atoms in two residue’s sidechains are within 3.5 angstroms, the typical maximum length of hydrogen bonds.

The ideal pseudoligand participates in a hydrogen bonding network that is not used to bind real ligands since such examples would be included in existing protein-ligand datasets. Besides typical small molecules that serve as ligands, we choose to exclude metal-coordinating residues even when buried, since these ions are typically not present in ligand-binding sites [53]. To exclude residues that participate in these binding sites, a strict filter is imposed that requires all atoms in a residue sidechain to be greater than ten angstroms away from any chemical group that is not water or a protein, effectively masking out residues near a small-molecule binding site.

Informed by literature indicating that multiple amino acids can coordinate to partially satisfy a polar residue [41] and that hydrogen bonding strength depends on rotamer configuration [42], we aim to maximize the number of residues that could potentially form a hydrogen bond to a candidate residue. Each residue is limited in the number of hydrogen bonding contacts it can make, because there are a limited number of donor and acceptor atoms on the sidechain group and because of space packing constraints around a residue. Arginine has the most electronegative atoms in its sidechain, 3 nitrogens, of any other amino acids, and experimentally we find that most residues have fewer than 4 amino acids that could serve as potential hydrogen bonding contacts. A score is thus generated to measure the richness of a hydrogen bonding neighborhood, normalizing this feature from 0 to 1 with a score of 0 assigned to residues with no contacts and a score of 1 assigned to residues with many contacts.
4.1.2 Burial

When a ligand docks into a binding site, polar residues within the binding site must release their bonds with free water or other polar amino acids and form more energetically favorable interactions with the ligand. This shares similarity with the environment of a polar residue found deep inside a protein structure [45], which must rely on internal hydrogen bonding networks to resist the need to bring free water into a hydrophobic region. The availability of solvent to a residue informs the degree to which it must rely on bonding networks, and this notion is captured by using a molecule’s solvent-accessible surface area, which measures the area of a molecule exposed to van der Waals contacts (Figure 4.1 B). Classically, this is calculated using the Lee & Richards algorithm [38], which defines the spherical volume around each atom, takes slices across these spheres, and integrates non-overlapping arc lengths to find the exposed surface area. We use the FreeSASA library [38] to calculate this for each candidate residue, as it takes as input a PDB file and uses atomic radii that correspond to specific amino acids and nucleic acids to provide additional specificity.

Before it can be used, a residue’s solvent-accessible surface area must be scaled so that it can be compared between amino acids. A bulkier sidechain for an amino acid leads to a larger surface area, and the sidechain rotamer can also affect these calculations because of the angles and distances between atoms in an amino acid. To correct for these effects, we measure a baseline solvent-exposed surface area for residues in a polypeptide chain by calculating it for the free amino acid in a tripeptide sequence, namely Gly-X-Gly where X is the amino acid of choice. Glycine, being the amino acid with the smallest sidechain, interferes the least with the exposed surface area of the residue in the middle of this tripeptide, and thus this solvent-exposed surface area can serve as the expected maximum level of exposure for that type of amino acid. Using this measurement to normalize the relative burial across amino acids, we calculate a score to measure the burial of a candidate residue normalized from 0 to 1, with a score of 0 representing the maximum exposure of the residue to solvent and 1 representing a completely buried, and thus solvent-inaccessible, residue.
4.1.3 Conservation and Diversity

Looking at a single residue and hydrogen bonding neighborhood in isolation can only inform us about how that network formed. Using the residue as a model for a ligand, we are particularly interested in how varied ligand binding sites can be developed around that chemical group, with the understanding that evolutionary changes to the overall protein structure should be represented and compensated for in alternative hydrogen bonding networks [54, 55]. Multiple sequence alignments (MSAs) offer a method to align a set of query sequences based on inferred homology. MSAs are used in this context to find ligand and ligand binding site pairings that are evolutionarily conserved and to calculate a heuristic for the diversity of a candidate pseudoligand’s hydrogen bonding neighborhood.

Multiple Sequence Alignments

We use multiple sequence alignments from OpenFold training data [56], which consists of MSAs for 140,000 proteins in the PDB and 16,000,000 UniClust30 clusters, grouping sequences from the UniProtKB database on the level of 30% pairwise sequence similarity. This threshold confers an MSA that has many matches to the sequence and has a large amount of diversity between sequences, and this depth and variety provide an ample field for the search for alternative hydrogen bonding neighborhoods. These files are stored in the a3m file format [57], which encodes alignments with insertions as lowercase letters, deletions (gaps) as dashes, and matches as uppercase letters. The nature of an MSA means that not every residue in the query sequence has a corresponding residue in the aligned sequences, and vice versa, as deletions and insertions occur.

To use both multiple sequence alignment data and structural PDB data, we must find a way to map between the residues indexed in a PDB file and the sequence used in the MSA. There are a few challenges to this approach, deriving from the fact that there is no explicit residue numbering in the MSA. The sequence used to construct an alignment may differ from the sequence found in the PDB file, either because additional amino acids were included in the experimental resolution of the protein structure or because amino acids in the sequence may be missing in the structural data because of resolution issues. Moreover, PDB files are arranged by bioassemblies, which may contain
multiple polypeptide chains, whereas MSAs are constructed based on a single query sequence. To tackle these challenges, we construct a pipeline to find the corresponding MSA column of a residue from its residue number in the relabeled PDB dataset, using metadata that tracks the mapping between the original PDB and newly reparsed files.

**Alternative Pseudoligand Neighborhoods**

Predicting where residues are spatially positioned from a sequence is a computationally intensive process since current approaches involve generating a predicted structure for each unresolved sequence. Instead, MSAs can be used as a heuristic for predicting which residues share a similar history and function. We are interested in finding pseudoligand residues that are conserved across the MSA, potentially indicating that the amino acid is engaging with a stable hydrogen bonding network, and we look for the converse in the neighborhood of that residue since we are interested in finding variations in the binding site around a pseudoligand that still form hydrogen bonding networks (Figure 4.1 C).

The conservation of a pseudoligand is measured as the percent similarity across amino acids found in each MSA column, indicating how much that residue has mutated across the various sequences. The variation of a column in the MSA is measured using two metrics. The first leverages the KL divergence between the distribution of amino acids found in the column of an MSA and the distribution of amino acids found across the PDB. The KL divergence can be interpreted as the expected difference between two distributions, thus capturing the distance between random substitutions in amino acids at a given position and selective conservation of amino acid identities [58]. For residues that have a high KL divergence, we expect to see that there is a bias away from the background distribution of amino acids driven by favorable interactions that select amino acids. Conversely, residues with a low KL divergence exhibit high variance and may participate in diverse contacting neighborhoods. To normalize this score across the PDB, we aggregate the KL-divergences calculated across all bioassemblies and identify the effective maximum divergence as four. Scores are scaled accordingly by giving residues with a KL-divergence of zero a score of 1, representing similarity to the desired uniform distribution, and residues with a KL-divergence of
four or higher a score of 0, representing residues that do not have high variance. Another measure of diversity is captured by the simpler metric of finding the percent difference across amino acids found in each MSA column, or the inverse of the similarity metric defined above. The two metrics capture different meanings behind the variance in the MSA and are used to find residues that participate in different hydrogen bonding configurations.

4.2 PDB Candidates Approach

Using the metrics described above, we begin the search and ranking of pseudoligand candidates with the Protein Data Bank (PDB), where we develop a pipeline to find good candidates from protein structure data and multiple sequence alignments. A scoring system is established based on the residue’s hydrogen bonding network and potential for alternative pseudoligand binding solutions and is designed to be extensible across future metrics of interest and parallelizable for high-throughput screening.

4.2.1 Processing PDB Data

The Protein Data Bank [36] contains an extensive collection of over 100,000 experimentally collected protein structures, containing information about their three-dimensional coordinates and metadata relevant to structure determination. The size and history of the PDB present challenges for the quality and formatting of this data, as earlier deposited structures are typically stored in the PDB file format while newer structures are stored in the mmCIF file format [59]. While the mmCIF format is a current, widely adopted standard, there are compatibility issues with legacy software that only supports the PDB format and direct translation between formats is not always possible due to domain restrictions in the PDB file structure. In addition, structures deposited in the PDB are formatted differently depending on the author, varying across residue index labels and the dividing of the protein complex across chains and segments. A standardized way to ingest this data is thus necessary for both tractability and so information extracted from these files are compatible with the pseudoligand search pipeline.
To standardize these files, we leverage a parsing script made by Benjamin Fry to convert mmCIF-formatted files from the PDB into relabeled PDB files, relabeling residues within each biological assembly so that chain, segment, and residue numbers are consecutively ordered. For mmCIF files that contain more chains than can be expressed in the PDB file format, the atoms belonging to the least significant chains in the mmCIF file are grouped into one new chain in the PDB file and accompanied with explicit metadata that tracks segments and residue numbers to distinguish between the chains; this generally does not present an issue as such atoms often encode small molecules rather than the protein structures needed for pseudoligand candidates of focus. Additional filtering is done to exclude deposited structures that are low-resolution (defined as structures with less than 3.5-angstrom resolution or resolved using methods other than x-ray crystallography or cryo-EM), sequences with over 10,000 residues for computational efficiency, and structures deposited after 06/01/2022 for reproducibility.

### 4.3 AlphaFoldDB Candidates Approach

Augmenting existing protein-ligand datasets with examples from the PDB is meaningful but limited by the number of resolved structures available. While the PDB contains around 100,000 entries [36], sequence databases like UniProtKB contain over 250 million sequences, of which 500,000 have been reviewed for quality and are manually annotated as part of the Swiss-Prot set [60]. To extend our approach of finding pseudoligands within the PDB, we present a method for leveraging computationally predicted protein structures to further build a pseudoligand candidate database. While there are also limitations to using computationally predicted structures, they provide a scalable solution for finding hydrogen bonding networks for not-yet-resolved sequences, exploit the strengths of current protein structure prediction models, and serve as a bridge between advances in protein folding and that of sequence prediction. We leverage multiple sequence alignments as a heuristic for finding diverse residue neighborhoods and use the AlphaFold Protein Structure Database (AlphaFoldDB) to validate our approach using precomputed structures.
4.3.1 Computational Bootstrapping

The challenges associated with the experimental structural determination of proteins stem from the physical limitations of current physical methods. This is an issue not only of the scale and speed at which researchers can resolve such structures but also of coverage and inclusion. For instance, transmembrane proteins have hydrophobic residues on their surface that make it difficult to express, purify, and crystallize these proteins [61]. These proteins are vastly underrepresented in the PDB but serve as targets for more than half of US FDA-approved drugs. This motivates the use of non-experimentally resolved structure data to address gaps in the training data.

The incorporation of synthetic data into training must be done with caution. Machine learning models, when trained on generated data, can suffer from model collapse [62], a phenomenon where the model’s understanding shifts away from the data distribution underlying the original content due to compression or distortion from the generation process. The model collapse has been shown on variational autoencoders, gaussian mixture models, and large language models, but all generative models can suffer because of recursive training [62]. In the context of protein sequence prediction, a performance decrease in predictions may occur if training datasets do not filter out results generated by models also trained on synthetic datasets. However, single-pass inclusions of synthetic features have been shown to improve prediction performance when they augment original protein training data [63]. An approach involving a more limited inclusion of synthetic data focused on the strengths of a given model architecture and the confidence of the prediction itself may be helpful for dataset augmentation.

The sourcing of synthetic data used in this research is driven by the availability of compute resources and fit for the training task. We have access to a database of over two hundred million AlphaFold-predicted proteins in the AlphaFold DB [64], supplementing structures that are not available in the PDB and using it to predict structures of sequences found in the UniProtKB sequence database. Our focus on hydrogen bonding networks necessitates that the positioning of residues within a predicted protein structure is close to what would be found experimentally, and AlphaFold predictions excel by having low RMSD values from experimental models, with a median distance of 0.96 angstroms [8]. This confers greater reliability upon the contacts we calculate in
our methods and more trust in the local structure of a residue’s neighborhood. Previous work on training sequence design models on structures predicted using AlphaFold2 indicates higher sequence recovery for buried residues compared to the sequence overall [65]. The model architecture also incorporates information from multiple sequence alignments and pairwise representations of amino acids in the sequence to get evolutionarily informed structures [8]. This improves the compatibility of our approach which uses the MSA to find structurally similar neighborhoods for buried pseudoligand residues.

We address the challenges of AlphaFolded structures with a careful selection of residues and neighborhoods to build upon. AlphaFold is trained on the task of predicting structures for single-chain proteins, and the AlphaFold DB structures do not contain protein complexes that are seen in the PDB bioassemblies [8]. Our focus on buried polar residues is compatible with these models, but it removes from consideration residues that facilitate interactions between protein chains which would otherwise be good candidates for hydrogen bonding networks. The effects of point mutations in sequences as well as the presence of non-protein modifications such as ions or ligands are also not explicitly available for AlphaFold predictions [65]. To avoid these pitfalls, we only select sequences from MSAs that have high homology to monomeric query sequences, increasingly the likelihood the UniProt is also monomeric, and exclude residues near binding sites or areas between protein chains to minimize discrepancies between AlphaFold predictions and the experimental structure.

4.3.2 Metrics Modifications

To allow for a direct comparison between candidate residues from the PDB and their corresponding residues in the AlphaFolded sequences, we carry over the same metrics by measuring the burial of a residue and its neighborhood, the conservation and diversity of a residue and its neighborhood, and the number of hydrogen bonding contacts between sidechains. When calculating a score depending on the MSA, such as the diversity of a neighborhood, we exclude from consideration contacting neighbors that do not have a position on the MSA, as there is no alignment information about that residue relative to the original query sequence.

In addition, we confer information about the confidence of the placement of a residue within
computationally generated structures by measuring the predicted local distance difference test (pLDDT) of a residue and its neighborhood, which is an AlphaFold score assigned to each residue that estimates the degree to which the coordinates are accurate [8]. The pLDDT metric ranges from 0 to 100 and scores above 90 are considered accurate to the rotamer level, which is necessary for our consideration of hydrogen bonding between sidechains. Segments of the structure that have low confidence are masked out to avoid confusing the model with poor training data.

### 4.4 Scoring and Construction

A holistic scoring system is made that considers important metrics as defined above and aggregates them into a single value that can best measure how good a given residue would be as a pseudoligand for further analysis. A comprehensive search is conducted across all bioassemblies available in the PDB that have been parsed and relabeled according to Fry’s script, and we consider how many other residues the residue would be in hydrogen bonding contact with, the conservation of the residue in the MSA, the diversity of the residue’s neighborhood across the MSA, the burial of the residue, and the burial of the residue’s neighborhood. Each metric has been scaled from 0 to 1, with one being the most desirable, and a linear combination of these scores is used to generate an overall candidate score. We filter out irrelevant candidates from consideration using a few hard thresholds. We exclude residues that do not have any hydrogen bonding contacts or are participating in a binding site. We also require a similarity score of more than 50% within a column of the MSA to find highly conserved residues and a relative burial score of more than 50% to ensure that the residue does not have large solvent exposure.

Once a search for pseudoligand candidates in the PDB was completed, selected residues that had a high candidate score were used to find potential pseudoligand candidates in unresolved UniProt sequences using MSAs. A mapping between the residues in a structure from the PDB and the residues in an AlphaFolded structure was created by bridging the alignment of relabeled PDB indices and their location in the MSA query sequence with the alignment of the MSA hit sequence and the indices of the AlphaFolded structure. The A3M formatting of MSAs leaves the query
sequence, usually the primary sequence of a PDB entry, unchanged while sequences that follow are marked with insertions and deletions and may contain a fragment of the primary sequence or the full sequence with gap padding. As such, we must preserve the relative alignment of residues across the MSA as we determine the alignment of the residues from the MSA sequence and the primary sequence. We use this in conjunction with the mappings between our reparsed PDB files to establish a clear indexing directly between residues, which can be used to find corresponding residues across proteins that participate in similar hydrogen bonding networks based on sequence homology. Corresponding pseudoligands in the MSA sequences were found only for high-scoring candidates, meaning that the original residue in the PDB had a high likelihood of participating in interesting hydrogen bonding networks. The reasoning behind this approach is that the MSA can be used to point where similarly good candidates can be found in a related protein, so it is best to start with leads from empirical structures that mirror that for which we are searching.

The AlphaFold DB API was queried for sequences that had existing predicted structures, and these structures were downloaded as available as PDB files. Sequences in the MSA that did not have a previously predicted AlphaFold structure were skipped to improve efficiency and focus on high-throughput screening. These computationally generated structures are then reparsed to align the formatting with those of the reparsed PDB files, and a score is generated for each corresponding residue to verify how similar it is to a good pseudoligand candidate. At the end of this pipeline, a database has been constructed with protein sequences and structures from the PDB and AlphaFoldDB, with each residue labeled with a score demarcating how well it can be used as a pseudoligand example.

4.5 Characterizing the Pseudoligand Candidate Database

To understand the types of residues selected as pseudoligands, we qualify the database across experimental and computationally predicted structures and measure how scores for pseudoligand viability are distributed. Because of prefiltering cutoffs used to determine which residues could be pseudoligands, which included a requirement of partial burial and partial conservation, selected
Figure 4.2: Distribution of Pseudoligand Metrics. A) Pseudoligand scoring is relatively normal, and AlphaFoldDB pseudoligand scores are higher because of the MSA selection process. B) Pseudoligand residues skew toward more polar amino acids and are rarely nonpolar. C) Distributions of the metrics that constitute the candidate score are not normal but indicate that prefilter conditions lead to good pseudoligand examples that participate in diverse and involved contacting neighborhoods.
candidate residues are more likely to exhibit desirable pseudoligand qualities, as indicated by the skew toward one in the scores of many constituent metrics that make up the candidate score. Pseudoligand candidate scores generated for structures from both the PDB and the AlphaFoldDB (Figure 4.2 A) are close to normally distributed, even as the calculated scores for each metric that goes into that final score are mostly not normal. These scores are also roughly distributed around the expected mean since PDB scores are centered around 0.5 and AlphaFoldDB scores are centered around 0.8, which stems from a search of AlphaFoldDB pseudoligands using MSAs from PDB pseudoligands that had a score of 0.8 or higher. The residue identity across both data sources (Figure 4.2 B) also indicates that pseudoligand residues tend to be polar amino acids, and amino acids that are positively or negatively charged are more heavily represented than polar but uncharged amino acids, which matches expectations around the strength of hydrogen bonding networks formed and the amino acids that would need to participate in them to stabilize charges. Metrics measuring the burial of a pseudoligand residue and its neighborhood indicate that most residues found are heavily buried (Figure 4.2 C). This scales as the power distribution, which matches expectations about how residues are placed within a globular protein, and the distribution of neighborhood burial is more left-skewed because pseudoligands may be found near residues that block solvent exposure. Pseudoligand candidates were likely to be conserved across the MSA, and many of these residues participated in similar neighborhoods across their MSA, indicating that hydrogen bonding networks were firmly established across phylogeny. Differences between the conservation of PDB neighborhoods and AlphaFoldDB neighborhoods stem from the selection of corresponding residues, as high-scoring PDB pseudoligands are measured to have high diversity in neighborhoods, meaning that the constructed AlphaFoldDB examples also have higher diversity compared to the base frequency across all sequences. From the count of contacting neighbors across the PDB and AlphaFoldDB, we see again that naturally, having multi-residue hydrogen bonding networks is rare but that the selection for highly involved residues during the search for pseudoligands within AlphaFoldDB led to a heavier selection for residues that participated in networks (Figure 4.2 C).

We also quantify the size of the pseudoligand candidate database to measure the scale of aug-
Structures and residues that can serve as high-quality pseudoligands are rare.

Table:

<table>
<thead>
<tr>
<th>Structures with Pseudoligands</th>
<th>Scored Pseudoligands</th>
<th>High Quality Pseudoligands</th>
</tr>
</thead>
<tbody>
<tr>
<td>PDB</td>
<td>65,754</td>
<td>684,286</td>
</tr>
<tr>
<td>AlphaFoldDB</td>
<td>45,578</td>
<td>113,978</td>
</tr>
</tbody>
</table>

Figure 4.3: **Breakdown of Pseudoligand Candidate Counts.** Structures and residues that can serve as high-quality pseudoligands are rare.

mentation provided, illustrated in Figure 4.3. Of the 261,205 PDB structures available, 65,754 of these structures contained residues that passed the filtering threshold of being a potential pseudoligand, and pseudoligand candidate scores were calculated for 684,286 residues. High-quality pseudoligand candidates, defined as residues that received a score of 0.8 or higher, were found across 8,214 structures with 10,733 residues. 45,578 structures from the AlphaFoldDB were added to the training data, with 113,978 residues assigned a pseudoligand candidate score and 23,710 of those residues being high-quality pseudoligand candidates. For pseudoligand candidates added from AlphaFoldDB, we are also interested in how many unique alternative solutions were added to the database, since our goal of including synthetic data was to find different residue networks interacting with the pseudoligand residue. To measure this, we count the number of pseudoligands that interact with at least one different residue compared to the neighborhood the pseudoligand interacts with in the PDB structure. We find that 23,072 unique examples are added to the database. Combined, the pseudoligand candidate database marks around thirty thousand residues that can serve as prime examples of pseudoligands, providing a meaningful addition to the over one hundred thousand ligand-binding structures that are available in the PDB [66].
Chapter 5

Training with Pseudoligands

In the previous two chapters, we proposed a novel search for examples of protein-ligand hydrogen bonding networks by modeling residue sidechains as interacting small-chemical groups. We applied this to experimental structural data from the PDB as well as computationally predicted structures from UniProt using AlphaFoldDB. Here, we describe and evaluate how pseudoligands can augment existing ligand binding databases, providing diverse training data for machine learning models tasked with predicting sequences that engage with hydrogen bonding. We use this pseudoligand candidate database in a machine-learning model, introducing preferential sampling of pseudoligand residues and training data with predicted structures containing pseudoligands. We then look at various metrics of machine learning performance to see how this database can be best leveraged.

5.1 LASErMPNN Architecture

We use the pseudoligand candidate database during the training of Ligand Aware Sequence Engineering Message Passing Neural Network (LASErMPNN), a graph neural network that predicts amino acid sequences that fold given an input protein backbone and a ligand to bind specifically to. LASErMPNN has three training phases. The first trains a ligand encoder module to build representations of the atoms within the ligand molecule, the second trains LASErMPNN on structural data from the PDB to learn sequence-structure and ligand-structure relationships, and the
third finetunes LASErMPNN specifically on ligand-protein binding in high-quality experimental protein structures. The dataset described in this thesis is incorporated into the second phase of LASErMPNN training to introduce a guided method of pseudoligand sampling and to expand the training data available to the model for learning small-molecule binding interactions and hydrogen bonding networks.

5.1.1 Objectives and Evaluation

LASErMPNN is training with the objective of minimizing loss across sequence recovery, MSA residue conservation, and chi-angle accuracy. To calculate sequence recovery, the cross entropy is calculated between the sequence logits output by the LASErMPNN prediction and the actual sequence for a given PDB file. This calculation provides a measure for how likely the actual sequence is to be drawn from the distribution described by the model prediction, and accurate, confident predictions are favored since they indicate that the model can place residues as expected within a sequence. The MSA residue conservation adds additional nuance to this measure of sequence recovery, motivated by the presence of alternative structures within real protein structures. This measure weights the cross-entropy loss penalty for predicting an incorrect residue at a given index in the sequence based on the frequency of amino acid substitution at that index according to the multiple sequence alignment. As such, the model is penalized less for predictions that are plausible in the MSA and more for predicted residues that are not conserved or are not present within the protein’s evolutionary history. This loss term makes concrete the objective of making incorrect sequence recovery predictions gracefully, hopefully allowing the model to extend beyond the task of sequence recovery and toward informed sequence design. Lastly, the chi angle loss measures the cross entropy for the chi angles of the residues, helping the model predict both correct rotamers as well as the sequence itself. This objective is important for ensuring that hydrogen bonding residues can make contacts, as the placement and direction of residue sidechains are important criteria for the strength of a hydrogen bonding network.

Evaluation of model predictions leverages a few more nuanced metrics to measure the specific performance of ligand-binding protein design. Residue predictions that are made correctly are
compared to the placement of that residue in the ground truth coordinates by calculating the root-mean-square deviation of the sidechain rotamers. The construction of rotamers from the model-predicted chi angles allows for a granular atomic-level comparison of the prediction and helps measure the cumulative accuracy of chi angles. We also measure the accuracy of first shell sequence predictions, defined as residues that are in contact with ligands. This metric helps gauge the performance of binding site design as these sites rely on specific chemical groups in the residues to form a favorable environment for ligands.

5.1.2 Model Tuning

We pay close attention to the construction of the training and test dataset available to LASErMPNN to minimize data contamination. Because protein prediction models are often trained on large databases such as the PDB, contamination may occur if sequences and structures that are similar in homology are included during training. To minimize this effect during test evaluation, we hold out the streptavidin class of proteins which is chosen because it features a binding site that has one of the strongest ligand interaction environments found in nature [67]. The protein class features a binding site that is incredibly specific and high affinity, which offers a good benchmark for the model’s ability to design residue binding sites. In addition to removing streptavidin proteins from training, we also preprocess the data so that proteins with an MSA cluster similarity to streptavidin of greater than 30% are excluded. This train/test split ensures that the model has minimal knowledge of binding sites like that of streptavidin and ensures that evaluation during testing is representative of the model’s ability to learn protein design principles instead of only reproducing what is seen in the PDB.

To reduce overfitting to the training data, LASErMPNN implements strategies at various steps of the training process to improve test performance. During training, noise is added to the coordinates of protein structures to model natural variance that occurs during protein folding. A small percentage of residues in the protein complex are also dropped out, allowing the model to selectively ignore information from residue nodes during the message passing update of the neural network. This method regularizes the model and helps LASErMPNN generalize beyond the training data by
preventing the layers in the neural network from coadapting and correcting prediction errors that would only apply to the data the model is exposed to during training. In addition, the decoding order of residues within the sequence is randomly generated so that the model can learn how to predict sequences with various sequence configurations, not just sequence orders that are found in nature.

5.1.3 Psuedoligand Candidate Database Sampling

LASErMPNN uses the strategy of sampling residues within proteins to be used as additional training data for ligand-binding interactions. It does so by isolating a selected residue, dropping out residues adjacent in sequence from the graph representation of the protein, and modifying the residue by adding methyl-amide caps to represent the backbone. The sampled pseudoligand thus becomes separate from the protein structure graph and its representation matches that of natural small-molecule ligands. There is at most one pseudoligand sampled per 250 residues in the amino acid sequence, and coordinates in the protein structure are K-means clustered so that pseudoligands are sampled randomly across the length of the sequence. The original approach of LASErMPNN randomly samples residues that are valid for pseudoligand pretraining, defined as residues that have at least three other residues whose heavy atoms are within 5 Angstroms of the selected residue’s sidechain. This sampling is normalized over amino acid identities to not oversample from more common amino acids.

The establishment of the scoring system for candidate residues provides a justified way to choose which residues in the structure are chosen to serve as a ligand, guiding training toward examples that are more likely to participate in stabilizing hydrogen bonding networks and other interactions that mirror ligand-protein binding sites. This biases the attention of the model during training toward difficult-to-design hydrogen bonding networks, deemphasizing the role of hydrophobic packing during sequence construction. We implement a strategy for pseudoligand selection that oversamples from pseudoligand candidates that received a high score in the database, choosing to include as many pseudoligand candidates that received a score higher than 0.8 as possible and falling back on random residue selection only when no other residues have been marked as good candidates for
pseudoligand selection.

5.2 Model Performance

We compare LASerMPNN performance across two training datasets, one with the original PDB structures and random pseudoligand sampling and the other augmented with structures from AlphaFoldDB and with prioritized pseudoligand candidate sampling. The training loss for both datasets (Figure 5.1 A) fared similarly across the three components of sequence, MSA, and chi angle loss, indicating that the inclusion of synthetic data from AlphaFoldDB did not hurt the model’s ability to predict protein sequence and rotamer configurations. The learning rates between the two models are also remarkably similar (Figure 5.1 B). A comparison of accuracy for sequence recovery overall and sequence recovery within first shell contacts of the ligand (Figure 5.1 C) also did not differ from the change in databases and suggests that augmenting with AlphaFolded structures does not improve sequence recovery. These trends may be explained by the type of information that is encoded in the computationally generated structures since AlphaFold may reencode structural information already available in the PDB. As a result, for metrics that measure general sequence and structural prediction performance, the model should not see a significant increase in performance.

Looking at the residues predicted by LASerMPNN during inference may provide a better view of what types of hydrogen bonding networks are established. We compare predictions on PDB entry 4JNJ, a held-out structure of streptavidin, using the base model weights and the pseudoligand candidate database model weights. From Figure 5.1 D, the ligand in the PDB structure makes seven hydrogen bonds to other residues in the protein. While both models make similar numbers of hydrogen bonds, the model trained on the pseudoligand candidate database builds a network with residues that match more similarly to the original PDB structure both in terms of the residues that interact with the ligand and the rotamer positioning of the residues. Notably, in the base model, a large sidechain is added to bind to the carboxyl group on the left, introduces a different amino acid to interact with the rings on the bottom, and introduces an aromatic interaction instead of forming more hydrogen bonds with the ring. This case indicates that LASerMPNN trained with
Figure 5.1: LASERMPNN Performance with Pseudoligand Candidates. A) Test losses decrease for models on both datasets B) Training rates are remarkably similar between both datasets C) Test sequence recovery on the global and first shell level achieve over 50% accuracy D) Residue networks designed by LASERMPNN trained on the pseudoligand candidate database mirror the networks found in the PDB more closely than the base model.
the pseudoligand candidate database can make more specific predictions that prioritize hydrogen bonding configurations between ligands and the surrounding residues.
Chapter 6

Conclusion and Future Work

Through this thesis, we investigated how to improve the performance of ligand-binding protein sequence design models using augmented training data focused on hydrogen bonding networks. We began with a discussion about characteristics important to protein folding and the specific challenges surrounding the design of small-molecule binding proteins. We then investigated the importance of hydrogen bonding networks in creating stable structures and tight ligand binding, illustrating how augmenting model training data with additional examples of hydrogen bonding networks and biasing predictions toward incorporating hydrogen bonds can improve model performance. Our main contribution is the creation of a pseudoligand candidate database that can be used by machine learning models to train on sequences and structures that contain residues interacting with diverse hydrogen bonding networks. This database provides a scoring system that ranks residues by how well they can be used as an example of ligand binding, allowing granular control over how much data is incorporated and how pseudoligands are sampled during training. By training LASErMPNN, a graph neural network, on the pseudoligand candidate database, we demonstrate that preferential sampling of pseudoligand residues does not harm model performance for overall sequence recovery, chi angle prediction, or MSA substitutions and that binding site design can demonstrate a closer similarity to experimental sequences compared to the base model. The formation of this database provides an additional resource and pipeline for the expansion of protein-ligand examples and the incorporation of synthetic data into protein sequence design models.
A key component of the pseudoligand candidate database is the development of a scoring system to measure how well a residue can serve as a pseudoligand and the processing infrastructure for the search and addition of pseudoligands. For instance, as structures are deposited into the PDB and as computationally predicted structures become more accurate, the database can be further expanded to find additional pseudoligands and buried hydrogen bonding networks. The availability of these scored residues provides an important starting point for the discovery of higher-order interactions used in ligand-binding proteins, the preferential biasing of models toward building hydrogen bonding networks, and the discovery of diverse structures using homology. We hope that these strategies of data augmentation and pseudoligand sampling can inform future work in sequence design predictions. Our finding could be applied in alternative machine learning architectures and to train toward different neural network objectives including hydrogen bonding complexity and binding site specificity.

For many machine learning models used in the protein design field, higher quality training data is still necessary for the improvement of model performance overall and a deeper understanding of the sequence space available for de novo designs. An additional consideration for the construction of robust binding networks is the role of second-shell interactions, which consist of the residues that interact with the binding site residues. Constructing hydrogen bonding networks requires stability at the binding site as well as support from the rest of the protein, and it remains an important challenge to understand how these wider interactions contribute to protein folding and stability. Having a deeper understanding of these involved networks can provide additional examples of favorable interactions that should be oversampled during the training of a neural network. A future direction of work is the expansion of the pseudoligand candidate database to take into consideration additional metrics that define ligand-protein interactions and to explore ways of including pseudoligands during training, either by varying methods of sampling or the construction of more varied pseudoligand chemical groups. Investigating how these measures can improve the quality of data labelled in the database as well as how the database can be most effectively used in training is an exciting avenue for improving model performance using this data strategy. Ultimately, the development of more accurate protein sequence design models depends on our ability to create
training regimens and datasets that draw upon the many complex interactions driving protein folding and function, and developments advancing this understanding of protein sequences and structures have the potential to be impactful and usher in future revolutions in protein design.
Bibliography


