Rotationally equivariant learning of generalizable protein structure-to-function maps

Michael Pun

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2023

Reading Committee:

Armita Nourmohammad, Chair

Philip Bradley

Miguel Morales

Program Authorized to Offer Degree:
Physics

University of Washington

Abstract

Rotationally equivariant learning of generalizable protein structure-to-function maps

Michael Pun

Chair of the Supervisory Committee:
Armita Nourmohammad
Department of Physics

Proteins play a central role in biology from immune recognition to brain activity. Although major advances in machine learning have improved our ability to predict protein structure from sequence, determining protein function from structure remains a major challenge. While the challenge of data availability has recently been alleviated due to computational structure prediction methods, the three-dimensional nature of protein structures complicates the application of traditional machine learning methods. Geometric deep learning offers a principled framework for efficiently extracting information from data which naturally respect physical symmetries. These symmetry-aware models have been shown to outperform and generalize better than non-geometric models. The goal of this thesis is to develop a minimal rotationally equivariant model to analyze local protein structures and systematically test its ability to generalize to relevant tasks in protein science. Here we develop Holographic Convolutional Neural Network (H-CNN), a rotationally equivariant neural network for predicting amino acid propensity based on local atomic micro-environments. We show that H-CNN's predictions quantitatively reflect the physical and chemical nature of amino acids leading to interpretation of H-CNN as an effective potential for amino acids. Subsequently, we use this interpretation to demonstrate H-CNN's generalizability in the zero-shot prediction of experimentally measured free energies of protein stability and binding. Finally, we apply H-CNN to the problem of determining T Cell Receptor (TCR) specificity by attempting to classify, predict, and design peptides that bind to given TCRs.

TABLE OF CONTENTS

Page

LIST OF FIGURES

GLOSSARY

ACE2: Angiotensin-converting enzyme 2 is the human protein that SARS-CoV-2 uses to gain access to human cells.

ANTIGEN: A molecule that is recognized by an immune receptor. This can be a protein, peptide, or other biomolecule.

CDR: Complementarity determining region. These regions of an immune receptor predominantly dictates the ability of the receptor to bind to and recognize different epitopes. These regions are defined as subsequences that usually form loop in the tertiary structure of the proteins.

EPITOPE: The particular part of an antigen that an immune receptor recognizes. The epitopes that T cells recognize are short peptides presented on the surface of cells.

HLA: Human leukocyte antigen. The name sometimes used to refer to MHC molecules in humans.

MHC: Major histocompatibility complex. A cellular surface protein found on all nucleated cells that displays peptide fragments of internal proteins.

pMHC: An abbreviation for referring to the peptide-MHC complex.

RBD: The receptor binding domain of SARS-CoV-2 is viral protein that CoV-2 uses to bind ACE2 and access human cells.

SASA: Solvent accessible surface area. The surface area of an atom, residue, or protein that is accessible to solvent.

SIDE CHAIN: The name used to refer to the R group of amino acids when they are used as residues in a polypeptide chain. The side chain is unique to each amino acid and determines the functional properties of that amino acid. Atoms in the side chain are named according to the Greek alphabet.

TCR: T cell receptor. An immune receptor found on T cells that is used by these cells to interrogate nucleated cells and detect infected pathogenic cells.

ACKNOWLEDGMENTS

First and foremost, I would like to thank my advisor Armita Nourmohammad for the opportunity to pursue this work and support throughout. I will always be grateful for your encouragement to embark on this project. It has been a tremendous and rewarding experience that would not have been possible without your scientific guidance, practical advice, and general encouragement. In my time in your lab, I have always appreciated your ability to guide and develop the members of your lab, and I feel very fortunate to be a recipient of your mentorship in my development as a scientist.

I would also like to thank all the members of the Nourmohammad group who have joined me in this project. Thank you, Jakub, for your guidance and advice. Thank you, Andrew, for sharing your trade with me at the beginning. Thank you, Colin and Assya, for sharing your wealth of knowledge and experience with me. Thanks to Quinn, Uchenna, Utheri, William, Eric, Ella, and Kevin for the opportunity to share this project with you all. Thank you, Arman and Gian Marco, for your collaborations on the extensions and off-shoots of this project. A special thanks to Zach, for your constant helpfulness as a labmate and a friend. Thank you for your patience in discussing problems, delicious lunches, and for your help to the end in your comments on this text. Also, thank you to the other members lab for your discussions and friendships including Giulio, Obinna, and especially Oskar for your helpfulness when I arrived in Göttingen and your friendship beyond.

And finally I would like to thank family and loved ones for their constant support from near and far. Thank you, Mom, for gifting me a love of learning, and, Dad, for your always useful pragmatic perspective. Thank you, Matt and Molly, for your care in always checking in. And thank you, Belinda, for being a constant source of fun, adventure, empathy, and love. I will always appreciate these characteristics, and I am extremely grateful you shared them with me during this time in my life.

Chapter 1

INTRODUCTION

Proteins are the machinery of life. They underlie the processes that drive living systems and are involved in almost all chemical reactions in biological organisms. Ion channels transport ions across neuron membranes to build the electrical potentials of cognition. Immune receptors protect vertebrates from harmful invaders by marking and neutralizing foreign pathogens. And replication of DNA is performed by an ensemble of proteins working together. These three key processes along with almost all sensing, signaling, transport, and storage in living systems involve proteins.

The importance of proteins in living systems is underscored by the sometimes devastating consequences of their malfunction. Misfolded proteins can form aggregates which in turn cause more misfolded proteins such as amyloid plaques in the brain which are closely linked to Alzheimer's disease [82]. Autoimmune diseases such as lupus and rheumatoid arthritis are caused by immune receptors that are triggered against self-proteins [161]. And famously more than 50% of human tumors are linked to mutation or deletion of one gene coding for the protein known as Tumor protein 53 (p53) [55, 133]. Remarkably, sometimes a malfunctioning protein can differ from a properly functioning one by just one mutation [61, 58]. Understanding any given protein's function is key to not only understanding life at a subcellular level but also to understanding disease and developing more potent medicines and interventions.

In the cases of both properly and improperly functioning proteins, their function (or lack thereof) is determined by physical interactions between the molecular building blocks of proteins, amino acids, and the atomic environment. Despite the importance of proteins in biology and medicine, how these amino acids exactly give rise to protein function is still not completely understood. There are three main factors that make studying proteins difficult: the diversity of protein functions, the complexity of protein physics and chemistry, and the

relatively small amounts of data at various scales of protein science.

While there is a good theoretical understanding of the fundamental interactions that govern atomic and molecular systems, first principles methods for determining how protein interact with cellular environments are prohibitively expensive to compute for an individual protein let alone many. Meanwhile, as the amount of data increases, data-driven machine learning (ML) models have come to the forefront of protein science due to their increasing ability to accurately predict experimental results coupled with their relatively cheap computational cost. The most notable example of this trend is AlphaFold's solution to the protein folding problem [66] which predicts a protein's structure from its sequence. Data-driven models like this are poised to guide the field of protein science in both scientific discovery and engineering efforts. In order to build robust structure-based models that can predict a protein's function accurately while still being sensitive to single amino acid mutations, an ML model should follow physical principles.

While symmetry are taken for granted in physical models, it has traditionally been neglected in structure-based models. Rather than utilizing the mathematical methods of physics, these models have borrowed from traditional ML algorithms rooted in computer vision, natural language processing, and network science. Only recently has the field of geometric deep learning united ML with physics in a principled and efficient framework to build physically sensible models which promise to generalize better on account of the inductive biases rooted in physical symmetries. It is an exciting time to be a part of this field with new methods being developed almost daily. It is my intention with this thesis to contribute to this field by developing a minimally rotationally equivariant neural network for determining amino acid propensities in local atomic environments and rigorously testing the generalization capabilities of such a network. Ultimately I will demonstrate that such a network can be used in a in a modular way to map structure to quantitative measures of protein function.

To motivate this work, in this introductory chapter, I will provide background on the three subject matters central to this thesis: proteins, ML, and symmetry. First in section 1.1, I will detail how proteins are structured, how this structure generally leads to protein function, and how the relationship between a protein's structure and function has been

a historically challenging problem. Then in section 1.2, I will detail how the field of ML offers new means to model structure-to-function maps from large datasets. At the same time, I will note some of the key challenges to building a robust ML model and how the recent marriage of physics and ML in the growing field of geometric deep learning offers solutions to some of the most common ML pitfalls. Lastly, in section 1.3, I will formalize the notion of symmetry mathematically and present the essentials necessary to understand the geometric deep learning model that is central to this thesis. These three background surveys of structural protein science, ML, and geometric deep learning will set the stage for my novel work on developing a data-driven rotationally equivariant structure-to-function map and applying it to common tasks and important systems in protein science.

1.1 Proteins

1.1.1 Amino Acids

The building blocks of proteins are molecules known as amino acids. The name *amino acid* comes from two functional groups (or sets of atoms that appear regularly and have similar properties across different molecules) that appear in each amino acid. They are the amino group, which is made up of one nitrogen and three hydrogens, and the carboxylic group, which is made up of one carbon and two oxygens and is an acidic group (see Fig. 1.1). The third functional group is known as the R group and is unique to each amino acid.

The 20 canonical amino acids are those which are encoded directly by the genetic code. This thesis focuses on these 20 canonical amino acids which are listed in table 1.1 and displayed in Fig. 1.2.

The atoms in an amino acid side chain are conventionally named following the Greek alphabet according to how far away they are from the central carbon which is denoted as the alpha carbon (often C_α). For example, in leucine the alpha carbon is followed by one beta carbon, one gamma carbon, and two delta carbons.

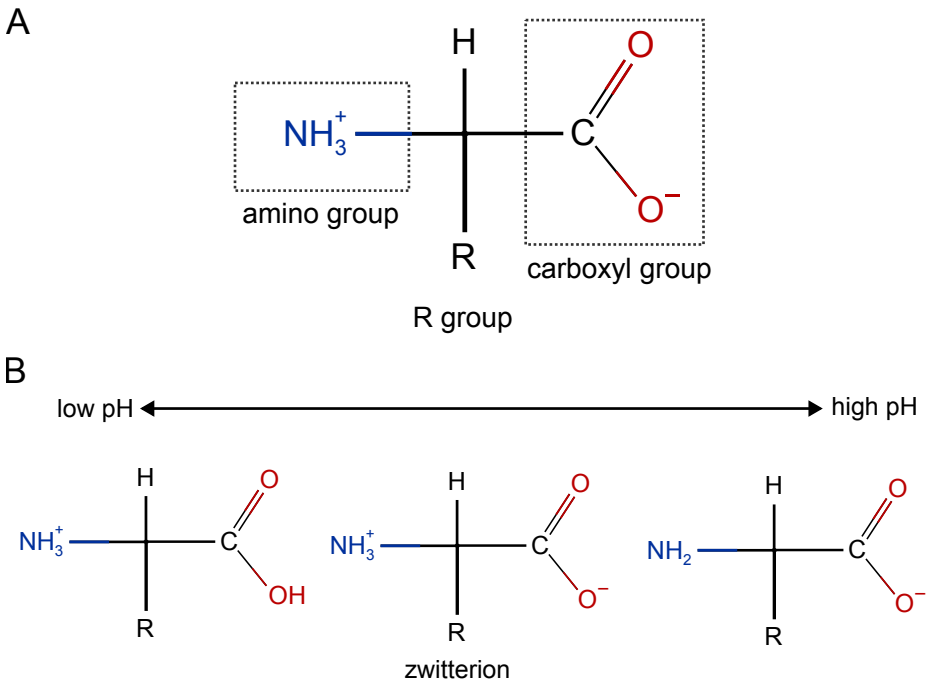


Figure 1.1: **Molecular structure of amino acids.** (A) The general structure of an amino acid is broken down into the carboxylic group, the amine group, and the R group or side chain. Note that because the amino group is basic and carboxyl group is acidic, each group will be oppositely charged at neutral pH and thus the molecule as a whole is termed a zwitterion. (B) As pH changes, the ionization of each side group changes at their respective acid dissociation constant.

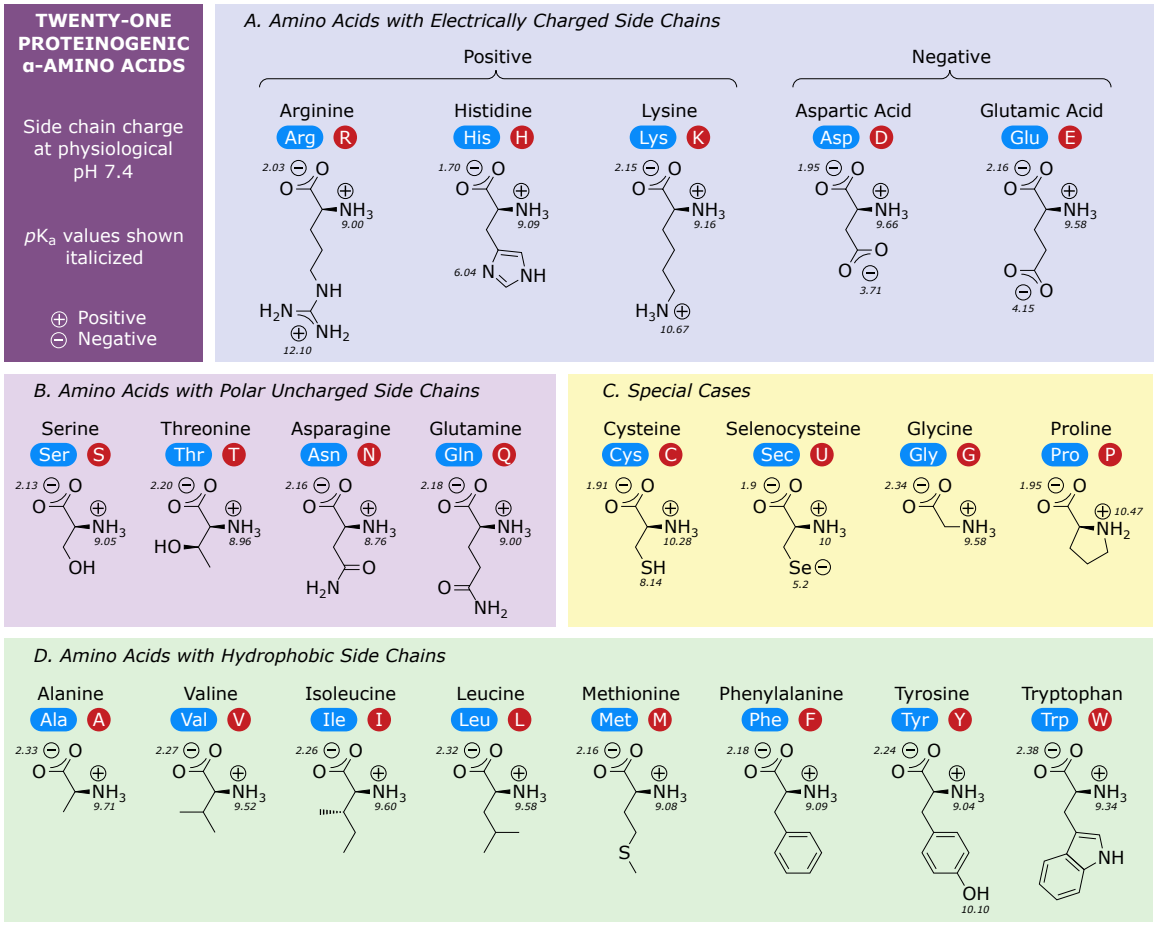


Figure 1.2: **The structures of the 20 canonical amino acids.**

Amino acid	Abbreviations	Mass (Da)	Surface area (Å^2)	pK_1	pK_2	pK_R
Glycine	G, Gly	57.05	75	2.2	9.8	-
Alanine	A, Ala	71.09	115	2.4	9.9	-
Serine	S, Ser	87.08	115	2.2	9.2	-
Proline	P, Pro	97.12	145	2.0	10.6	-
Valine	V, Val	99.14	155	2.3	9.7	-
Threonine	T, Thr	101.11	140	2.1	9.1	-
Cysteine	C, Cys	103.15	135	1.9	10.7	8.4
Leucine	L, Leu	113.16	170	2.3	9.7	-
Isoleucine	I, Ile	113.16	175	2.3	9.8	-
Asparagine	N, Asn	114.11	160	2.1	8.7	-
Aspartic Acid	D, Asp	115.09	150	2.0	9.9	3.9
Lysine	L, Lys	127.17	200	2.3	9.7	10.5
Glutamine	Q, Gln	128.12	180	2.2	9.1	-
Glutamic Acid	E, Glu	129.12	190	2.1	9.5	4.1
Methionine	M, Met	131.19	185	2.1	9.3	-
Histidine	H, His	137.14	195	1.8	9.3	6.0
Phenylalanine	F, Phe	147.18	210	2.2	9.3	-
Arginine	R, Arg	156.19	225	1.8	9.0	12.5
Tyrosine	Y, Tyr	163.18	230	2.2	9.2	10.5
Tryptophan	W, Trp	186.21	255	2.5	9.4	-

Table 1.1: **Summary of information associated with the twenty canonical amino acids.** Molecular mass is listed in Daltons (Da) which are defined as 1/12 of the mass of a neutral unbound carbon-12 atom. pK_1 and pK_2 are the acid dissociation constant of the carboxyl and amino groups respectively. pK_R is the acid dissociation constant of the R group if applicable.

Properties of amino acids

The atomic structure and composition of the R group determines the physicochemical properties of each amino acid (e.g., size, charge, polarity, etc.). When analyzing protein structure at the amino acid scale it is important to remember that amino acids have many orthogonal properties and their function in proteins can be related to any or all of these properties.

One of the most elementary properties of amino acid is size. The size of an amino acid can be described by the molecular mass or by the volume of space occupied by the molecule's atoms. By both measures, the smallest amino acid is glycine and the largest is tryptophan. Although size is a relatively straightforward property, sometimes a certain amino acid is key to the function of a protein simply because it is the only one that fits [121].

A more nuanced chemical property of amino acids is determined by if the amino acid acts as an acid or a base. Since the amino group is basic and the carboxylic group is acidic, the acid/base classification of amino acids is determined by the R group of the molecule. The acidic amino acids are glutamic acid and aspartic acid. The basic amino acids are arginine and lysine. Histidine is a weak base at physiological conditions ($pK_a = 6.04$) and can be considered in the basic category.

Since physiological conditions usually have a nearly neutral pH of 7.4, acidic and basic R groups will have a nonzero net charge. The distinction of charge then splits the amino acids along the same boundaries as the acidic/basic categorization. The positively charged amino acids are the basic arginine, lysine and histidine while the negatively charged amino acids are the acidic glutamic acid and aspartic acid. Charged amino acids often sit on the surface of proteins where they can interact favorably with polar solvent molecules via hydrogen bonds. Furthermore, charged amino acids are able to interact with each other via both hydrogen bonding and ionic bonding in a phenomenon known as a salt bridge. Studies engineering the charge of amino acids near active sites in enzymes have shown charge to change the pH of catalysis on the order of ~1 pH [136] or even change the free energy of binding to Ca^{2+} ions by ~7 kJ/mol [78].

One striking example of charge importance in protein structures is provided by the

Immunoglobulin G (IgG) antibodies. It is known that serum-derived proteins including IgG antibodies are negatively charged. However, under formulation conditions (pH 5-6), therapeutic monoclonal antibodies typically have a net positive charge [160]. This discrepancy is important to understand since it can lead to better understanding of which antibodies could have therapeutic functions. In this case and in general, determining protein charge is more complex than determining the charge of a single amino acid or even many amino acids. Rather the charge of a protein depends highly on temperature, pressure, salt concentration, salt type, and pH [160].

Neutrally charged amino acids can further be categorized by polarity. Polar amino acids contain either a hydroxyl group or simple amide group. The former category consists of serine and threonine while the latter is comprised of asparagine and glutamine. Charged and polar amino acids have favorable interactions with polar water molecules. Thus they are termed hydrophilic and are often present on the surface of protein structures.

Since hydrophilic amino acids readily form hydrogen bonds with water, they often are involved in determining specificity of protein-protein interactions. In fact, hydrogen bonds that are unsatisfied in complex can change the free energy of interaction by 0.5-6 kcal/mol [39, 16, 68]. One notable example of the importance of polar amino acids is in intrinsically disordered proteins (IDPs). IDPs are proteins that do not have a well defined structure except when they are interacting with partner proteins. Polar amino acids are found at higher frequencies in IDPs and are thought to account for more interactions in the complexed structure of IDPs than in ordered proteins [154].

In contrast to hydrophilicity, amino acids that are uncharged and apolar have unfavorable interactions with water and are hydrophobic. Their R groups often exhibit hydrophobic interactions with each other and are important for core stabilization. Hydrophobic interactions are important for protein stability and are thought to be a main driving force during the early stages of protein folding [97, 36].

While hydropathy is an important factor in proteins, solvent is usually not modeled explicitly in protein structures. Instead, the amount of solvent accessible surface area (SASA) is calculated to approximate the amount of potential for interaction an amino acid will have with water. Hydrophobic amino acids are expected to be buried in protein cores with no

SASA while hydrophilic amino acids are expected to be near the surface where their polar groups can interact with solvent.

Aromatic amino acids are those that contain a planar cyclic ring that has increased stability due to its ability to form π bonds. Aromatic amino acids are generally hydrophobic and are consequently often found in protein cores; however they also have been shown to play a role in protein-DNA binding [9, 54]. The aromatic amino acids are tryptophan, tyrosine, phenylalanine, and histidine.

Aliphatic amino acids have R groups composed of carbon atoms in open chains (i.e., not aromatic rings). Aliphatic amino acids are non-polar and hydrophobic. The aliphatic amino acids are generally considered to be alanine, isoleucine, leucine, proline, and valine. Despite the fact that its R group contains sulfur, methionine is sometimes considered with the group of aliphatic amino acids due to its generally unreactive and non-polar nature.

Overall, the unique R group of each amino acid endows it with a unique combination of these properties. A visualization of these properties and their overlapping and non-orthogonal nature is provided in Fig. 1.3. In any protein, one or more of these properties may contribute to the appearance of a specific amino acid at a site in the protein. Sometimes a hydrophobic amino acid is necessary and sometimes a small hydrophobic amino acid is necessary. In general, the importance of each property is not immediately evident from the appearance of a specific amino acid. Furthermore, how these properties contribute to the overall protein function is complicated by many factors including pH, temperature, pressure, salts, presence of other amino acids, and presence of solvent.

Peptide bonds

A property shared among all amino acids (and key to their role in proteins) is the ability to form a covalent bond between themselves. This bond is known as a peptide bond and involves the amino group of one amino acid with the carboxyl group of the other. Specifically the OH group on the carboxyl group and a hydrogen from the amine nitrogen dissociate from each amino acid and form water. In their place a covalent bond is formed between the carbonyl carbon and the amine nitrogen. This peptide bond can occur twice per amino acid,

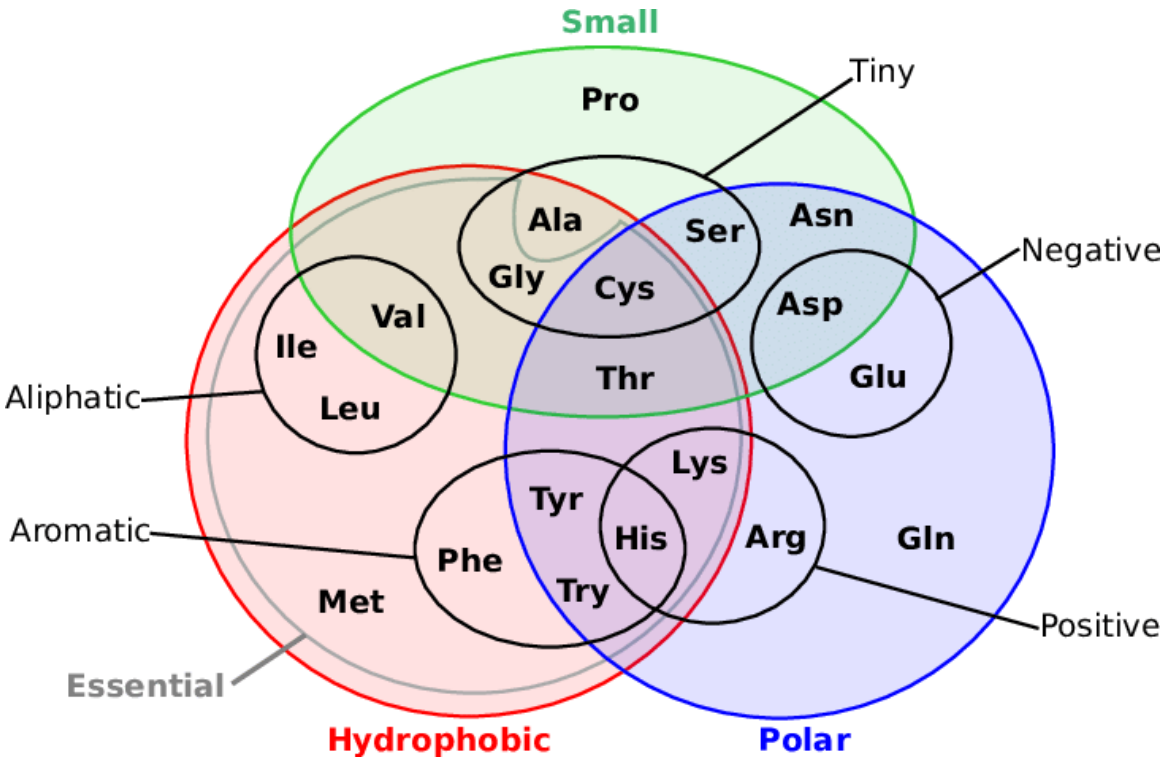


Figure 1.3: **Amino acid property map.** Amino acids are organized in this schematic according to their physical and chemical properties. This visualization emphasizes the overlapping and non-orthogonal nature of these properties. In proteins, the usage of a specific amino acid may depend on one or many of the properties of that amino acid.

once at the carboxyl group and once at the amino group. The capacity to form two bonds allows amino acids to form a molecular chain known as a polypeptide.

A polypeptide is said to have a backbone involving the amino and carbonyl groups with the R groups extending away from this backbone (see Fig. 1.4). In the context of a polypeptide, each amino acid is referred to as a residue and their R groups are consequently termed side chains. The structure of a polypeptide or protein can be classified at different levels of detail.

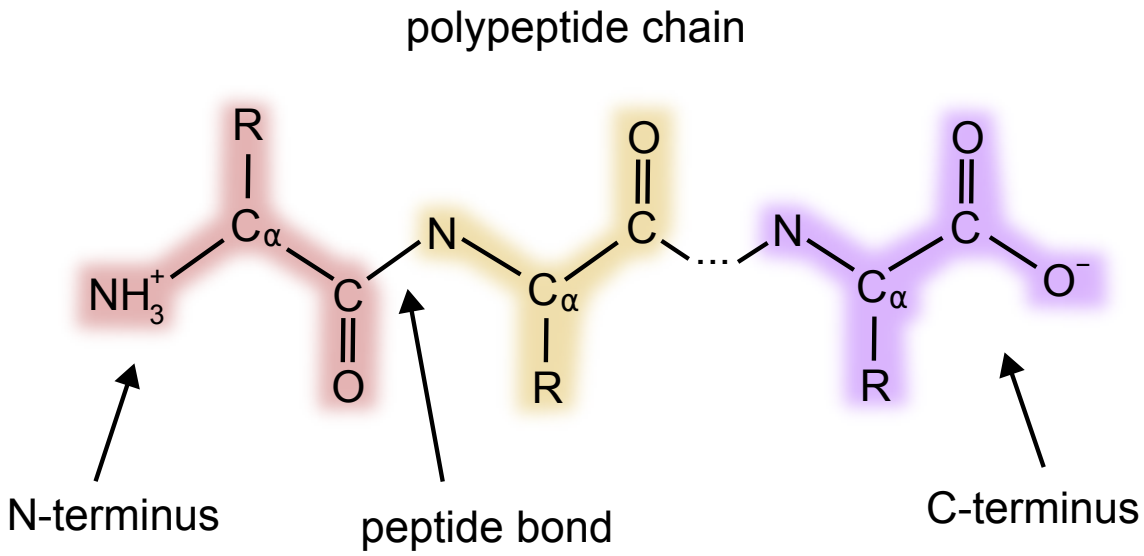


Figure 1.4: **Chemical structure of a polypeptide chain.** A peptide bond is shown between the first two amino acids in a polypeptide chain. Due to the fact that translation of proteins begins between the carboxyl group of the first residue with the amino group of the second, the beginning of the peptide is called the N-terminus. The C-terminus is shown at the opposite end of the polypeptide.

1.1.2 Protein Structure

Primary structure

The identity of a protein is determined by the sequence of amino acids that make up its polypeptide chain. This sequence is known as the primary structure of the protein and is encoded in the genetic code. The production of proteins from genes is famously described in the central dogma of molecular biology whereby DNA is transcribed into RNA which is then translated into proteins. In this process, different codons or units of three nucleotides encode for a specific amino acid. Since there are four nucleotides and 20 amino acids, it is clear that there is degeneracy as to which codons code for each of the amino acids. This degeneracy is out of the scope of this thesis, however a basic discussion can be found in [11, 98]. The process of assembling the polypeptide chain of a protein from the RNA code is known as translation. In natural proteins, the chain is usually initiated with a start codon which translates to a methionine residue. The subsequent amino acid's amine group is then bonded to the methionine's carboxylic group. This repeated process defines a direction in the sequence of a protein with the beginning called the N-terminus while the end is called the C-terminus (see Fig. 1.4).

Secondary structure

As a polypeptide is assembled during translation, the molecule "folds" into a three-dimensional (3D) structure according the interactions between the residues that have been translated. The folding of a protein is a spontaneous process that takes milliseconds for a typical protein [71]. In these 3D structures, the protein backbone often adopts regular structures that are notably stable due to hydrogen bonding networks between amine and carboxyl groups of the backbone. For example, alpha helices and beta sheets are two of the most common shapes and are highlighted in Fig. 1.6. These shapes are known as the secondary structure of the protein and can be determined from coordinates of the backbone atoms alone.

The geometry of a protein backbone is conventionally described by three dihedral angles at each site i noted by ϕ_i, ψ_i, and ω_i (see Fig. 1.5). ϕ_i is closer to the N-terminus of a residue

and is the dihedral angle between the planes defined by the two atomic trios $C_{i-1}N_iC_i^\alpha$ and $N_iC_i^\alpha C_i$. ψ_i is the dihedral angle between the planes defined by $N_iC_i^\alpha C_i$ and $C_i^\alpha C_i N_{i+1}$. ω is defined as the dihedral angle between the planes containing $C_{i-1}^\alpha C_{i-1}N_i$ and $C_{i-1}N_iC_i^\alpha$ (see Fig. 1.6).

Due to steric clashes (i.e., repulsive forces due to overlapping electron clouds of nearby atoms), ω is constrained to be approximately 180°. Steric clashes also limit ϕ and ψ, however, to a much lesser extent. Thus, secondary structures can be characterized completely by these two angles alone. These angles are often plotted against each other in the two-dimensional Ramachandran plot where ϕ is on the x-axis and ψ is on the y-axis (see Fig. 1.6).

Helices One of the most common categories of secondary structures are helices which are characterized by coils of the backbone. Alpha helices are the most common and involve hydrogen bonding between O_i and N_{i+4}; see Fig. 1.7. They are also localized in quadrant III of the Ramachandran plot; see Fig. 1.6. Other types of helices include the 3_{10} helix and the π helix in which O_i has a hydrogen bond with N_{i+3} and N_{i+5} respectively.

Strands and Sheets The other most common category of secondary structure are strands which are characterized by their general flatness. Beta strands are the most common type of strand with amine nitrogens and carbonyl oxygens extending in the plane of the strand alternatingly on each side; see Fig. 1.7. Beta strands exist in quadrants II and III in the Ramachandran plot; see Fig. 1.6. Due to their flat alternating placement of amine nitrogen and carbonyl oxygen, beta strands often form hydrogen bond in planar networks together and constitute larger forms of secondary structure known as beta sheets. These sheets are generally formed from parallel or antiparallel beta strands. Another sheet known as an alpha sheet also exists in protein structures; however it is much rarer than beta sheets.

Loops All backbone geometries that do not fall into the two previous categories are usually considered in the category of loops which have no well-defined secondary structure. Loops are often connecting other types of secondary structures and due to their flexibility can often play important roles in a protein's interactions with its environment. For example,

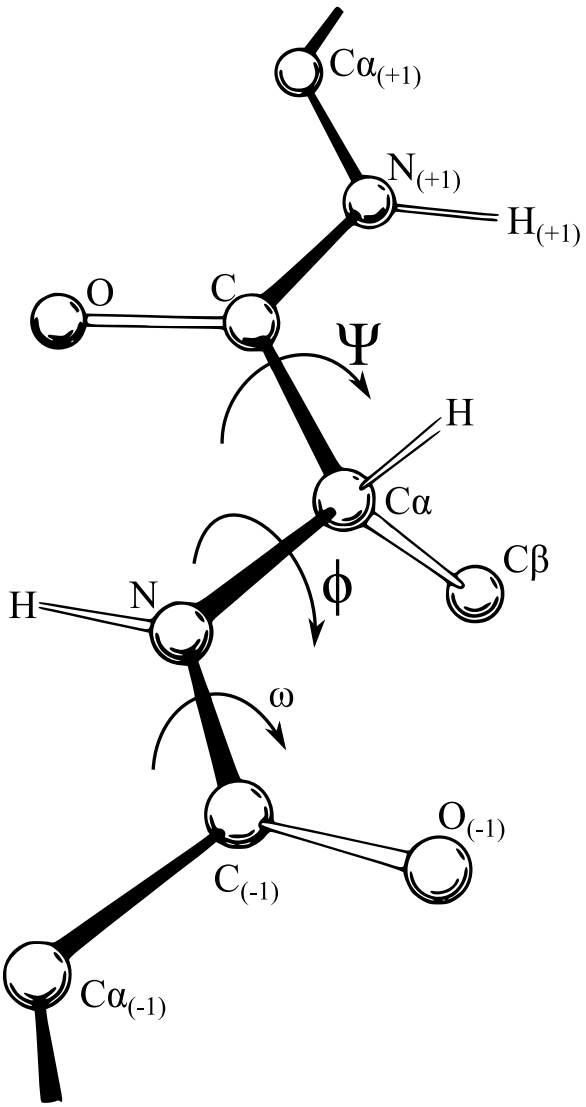


Figure 1.5: **Protein backbone and dihedral angles.**[†] A protein backbone is shown at residue i in a general protein. Side chains are generally hidden with the exception of the beta carbon at site i. Dihedral angles ψ, ϕ, and ω are shown around the bond shared between the three sets of plane-defining atoms.

[†] Image reproduced from wikimedia user Dcrjsr under Creative Commons Attribution 3.0 Unported license.

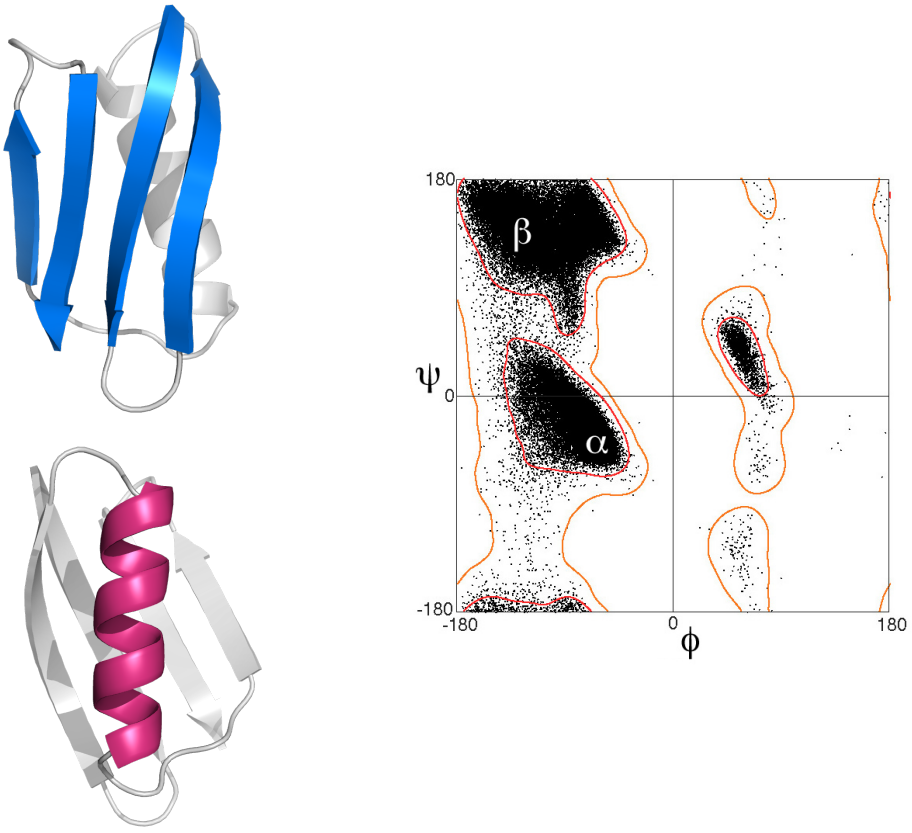


Figure 1.6: **Allowed backbone angles and the secondary structures they adopt.** Right is a Ramachandran plot of typically observed backbone angles [83].† Due to steric clashes only certain regions of the domain are allowed. The most populated regions are characterized by helices and sheets. The two most common type being alpha helices and beta sheets. Each of these structures is visualized at left.

† Image reproduced from wikimedia user Dcrjsr under Creative Commons Attribution 3.0 Unported license.

complementarity determining regions are integral loops in immune receptors that dictate their abilities to recognize pathogens. While loops are certainly less ordered than helices and sheets, recent work has shown that loops can be organized by structural properties [119] and these general structures can be useful for engineering purposes [57].

Tertiary structure

The tertiary structure of a protein is the 3D structure of the entire protein—side chains included. This structure is more complicated than secondary structure as it involves interactions between side chains and both other side chains and the polypeptide backbone. These interactions include hydrogen bonding, hydrophobic interactions, electrostatic interactions, and even covalent bonds in some cases.

A majority of experimentally determined protein structures are solved using X-ray crystallography while a minority are solved using NMR spectroscopy and electron microscopy. A review of experimental structural determination methods can be found in ref. [128]. Recently computational prediction of protein structure has been "solved" as AlphaFold2 [66] has been shown to be able to predict protein structure with high accuracy dramatically expanding the availability of accurate tertiary structures for proteins. This thesis will focus on utilizing tertiary structures to predict protein function.

Quaternary structure and beyond

Quaternary structure is generally the highest level of protein structure. Quaternary structure is given by the geometric arrangement of different protein chains in a larger complex that has some well-defined function. For example, quaternary structure of antibodies is described by the relative orientation of of two heavy chains and two light chains.

1.1.3 Protein Function

Once a protein is assembled in its 3D fold and complexes with any supporting domains, it can perform its intended function. Just as the structure of a protein is rooted in the interactions between amino acid side chains, the function of a protein is dictated by the nature of the

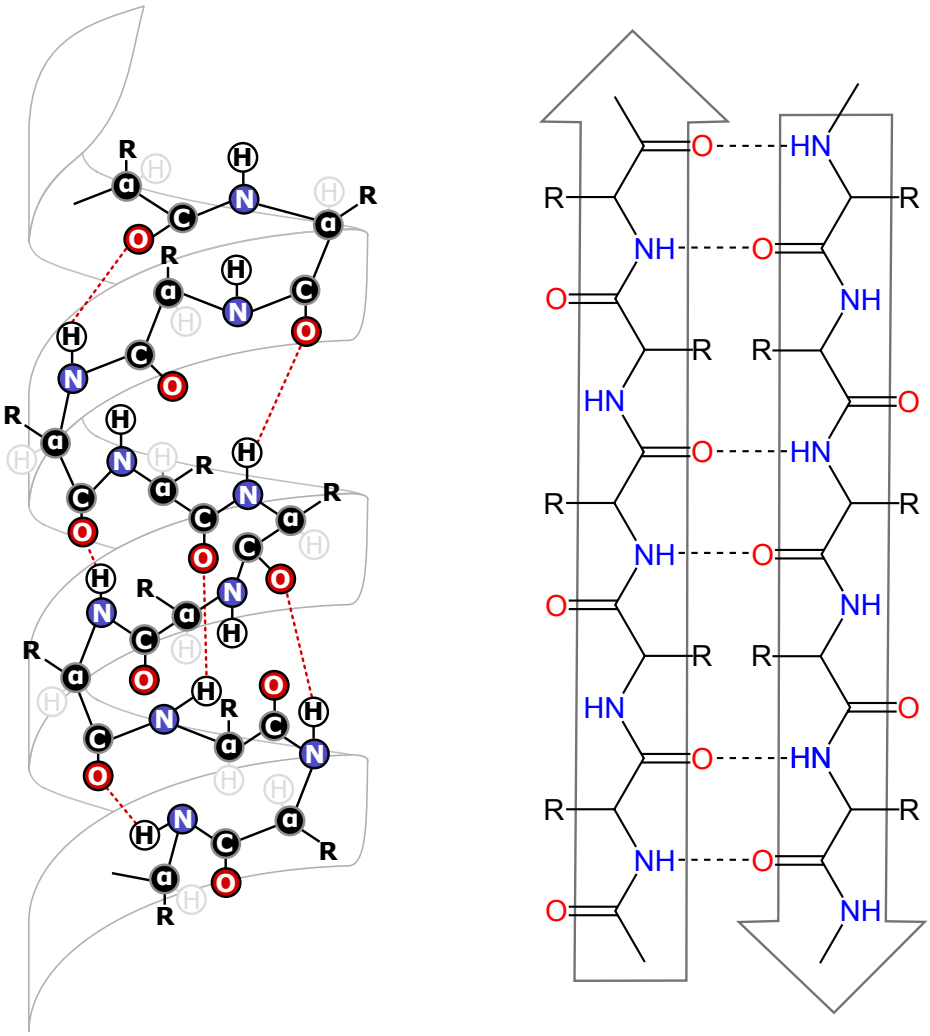


Figure 1.7: **Hydrogen bonding patterns in alpha helices and beta sheets.** On left, an alpha helix exhibits hydrogen bonding between the oxygen of residue i and the nitrogen of residue $i + 4$.[†] On right, beta sheets exhibit hydrogen bonding between antiparallel beta strands.

[†] Image reproduced from wikimedia user Danny Patrick Blair under Creative Commons Attribution 3.0 Unported license.

interactions between side chains and the surrounding molecular environment. Different proteins with different sequences and therefore different tertiary structures perform different functions. Proteins catalyze chemical reactions, synthesize and repair DNA, transport molecules, sense chemical signals, produce chemical signals of their own, provide structural support, and store molecules.

While these descriptions of protein function are qualitative, a protein's functions can also be described quantitatively by determining how well it performs its given role. For example, given two different immune receptors, one can quantify how efficient they are at neutralizing a given pathogen. While answers to the question "how well" are usually determined by a quantitative measurement of a protein's effect on its environment, theoretically such a quantity is still governed by the underlying interactions between a protein and its molecular environment. An antibody that binds strongly to a pathogen will generally neutralize it well. This thesis focuses on quantitative structure-to-function maps, so I will focus on surveying how a quantitative function such as binding affinity can be determined from the structure of a protein and its interaction partners.

Determination of function

Quantitative function is determined experimentally through the use of various techniques which correlate experimental observables with the underlying physical interactions. For example, in some attempts to measure protein-protein interactions, receptor proteins are attached to a thin metal plate that exhibits a phenomenon known as surface plasmon resonance. Solvated ligand proteins are then free to interact with the receptor proteins. Upon binding, the index of refraction near the sheet of the metal changes and directly influences the resonance of the surface plasmons. Thus binding can be directly inferred from observing the resonance. These types of experiments are low-throughput as they only screen one protein at a time. Other methods are higher-throughput, such as fluorescence activated cells sorting (FACS)-based methods that allow screening of many sequences simultaneously.

While experimental methods are generally accurate, they are expensive and time-consuming. Furthermore, some proteins are difficult to synthesize *in vitro* and are not

amenable to such experiments. The ability to circumvent experimental determination of function with reliably accurate computational methods would be a significant contribution to the field of protein science.

Complexity in computational function prediction

The main advantage experimental techniques have over computational methods is that they are accurate insofar as the experimental circumstances are controlled. Computational prediction of quantitative function must account for the complex mapping between protein identity and function that nature simply exhibits. For example, two proteins might have similar structures and similar functions as is the case with heme-binding proteins myoglobin. Specifically α globin and β globin, which have different sequences but similar structures, play similar roles binding the physiologically important heme group [152]. However, two proteins with similar sequences and structures may have disparate functions. For example, it is believed that the two murine proteins GDF11 and MSTN have distinct functions despite being homologs with 89% sequence similarity [132, 141] and a relatively small aligned root-mean-square deviation (RMSD) of atomic coordinates of 0.695 Å. Ultimately a robust and accurate function predictor must be able to simultaneously account for confounding situations like these. Prediction methods that take primary structure as input are called sequence-based and must account for both the map from sequence to structure and from structure to function. With the growing ability to computationally predict tertiary structures coupled with the simpler relationship between tertiary structure and function, structure-based methods offer an attractive avenue for prediction of quantitative function.

Bottom-up models

The most accurate method for computationally predicting protein function from structure is through the use of bottom-up models which are built from the ground up using first-principles. The most common bottom-up models are molecular dynamics (MD) simulations which accurately map physical interactions between heavy atoms as well as electron densities which are quantum in nature. While accurate, these computations are costly. In general,

the cost of MD simulations is proportional to the square number of atoms in the system making them intractable for large systems. Furthermore, MD simulations are often limited in time by numerical instabilities. Overall, the computational cost of these models makes them prohibitively expensive to apply to proteins.

Top-down models

A more efficient set of methods for computationally predicting function is the use of top-down models. Top-down models are models that are built using data-driven methods. Most of the computational cost of top-down models are front-loaded onto the training of the model, and once a good model has been learned, these models are efficient especially when compared with first-principles methods. Models for predicting quantitative function are generally divided by the nature of the molecular interactions that govern the function. Data-driven models have been developed to predict protein stability [114] and expression [126, 165], as well as for quantitative function prediction in the cases of binding [164, 14, 51, 140, 106, 81, 142, 148, 109, 162, 1, 24, 40] and enzymatic activity [75, 105, 53]. In general, these models fall under the umbrella of both statistical inference and ML. They range in complexity from linear regression to random forests to residual recurrent convolutional neural networks.

One key concern when building a top-down model is the quality of the dataset. In particular, data-driven models of protein stability prediction can be unreliable due to small biased datasets [114] and models of protein binding can also suffer from similar problems [68, 164]. This lack of generalizability stands in stark contrast to bottom-up physical models which are built to be applied to any atomic system generally.

Data-driven models also differ from first-principles models in their ability to rationalize about geometric data. While physical models respect symmetry by construction, data-driven models do not. Convolutional neural networks are one counterexample where translational symmetry is incorporated into a ML model, however a formal treatment of symmetry in ML models has traditionally been lacking. In this thesis, I will turn to the recently emerging field of geometric deep learning to build a symmetry-aware data-driven structure-based model of protein function. As will soon be shown, a proper treatment of symmetry also promises

to build more generalizable models. However, before we discuss geometric deep learning specifically, let us turn our attention to the fundamentals of ML to better understand the nuances of modern day data-driven models and how proper treatment of symmetry offers solutions to common problems in ML.

1.2 Machine Learning

Machine learning (ML) is the study of statistical algorithms that can learn models automatically from data. At the core of most ML methods are powerful function approximators that are trained to match the statistics of a dataset. These function approximators (often times in the form of a neural network) have two key properties. First, they are learnable which is usually ensured by employing matrix multiplications in linear layers. The weights in these matrices parameterize the model and these matrix multiplications are differentiable. Alongside a loss function that defines how well a model can match a dataset, this differentiability allows the model to learn automatically from the data by attempting to minimize this loss function. The second key property these approximators have is a high capacity for expressivity in the sense of being able to approximate a broad class of functions. This type of expressivity is ensured by the use of nonlinear activation functions. In fact, with enough capacity, simple neural networks can actually approximate all functions satisfying some criteria [29, 74] and are thus called universal approximators.

1.2.1 Bias and variance

While these two properties of learnability and expressivity, as well as a high quality dataset, are necessary to build a good ML model, training such a model is not as simple as just ensuring expressivity and differentiability. In fact, there are two main sources of error inherent to training any ML model: bias and variance. The bias in an ML model is the error due to space of accessible functions not including the true function that produced the data. For example, if one does linear regression on data that is cubic in nature, there will always be inherent error since the true relationship between the data is not expressible in the class of linear models. One can ensure low bias by using a very expressive model.

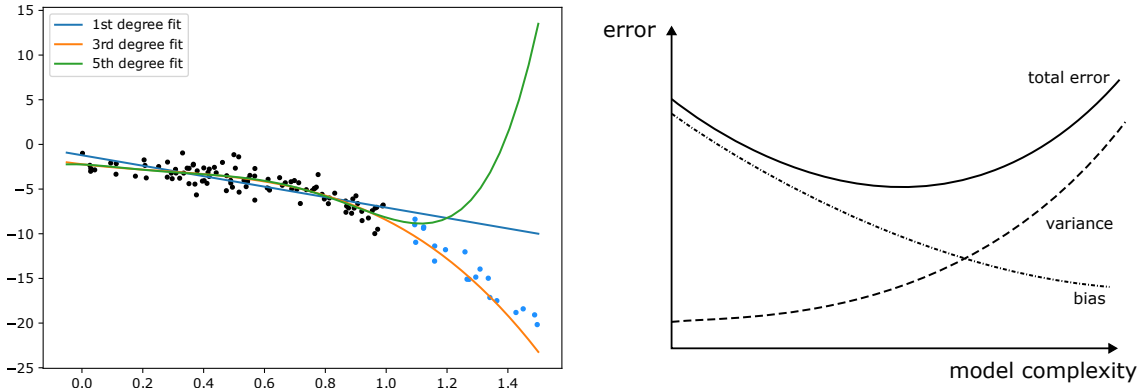


Figure 1.8: **Bias and variance illustrated in case of polynomial regression.** Noisy data is sampled from a fifth degree polynomial (scatter points). A subset of the data (black) is used to train three polynomial models of differing degrees under polynomial regression. The models are then compared to the data on held out samples (blue points). The linear model's errors exemplify bias since it is unable to capture the nonlinear nature of the data. Meanwhile the fifth degree model's error is due to variance. It has overfit on the training dataset and is unable to generalize to the new data. The third degree polynomial has both bias and variance but does better than the linear and fifth degree models at predicting the unseen data. This exemplifies the notion of an optimum in the bias-variance tradeoff shown at right.

Variance, the other source of error, is the error of the model due to the finiteness of the dataset. This error is the byproduct of the phenomenon known as overfitting and happens when an expressive model starts to learn the observed noise in a dataset. Since new data comes with different realizations of noise, overfitting will cause a seemingly well-trained model to make large errors on new data. Variance is decreased as models are less expressive since they have less capacity to fit the noise.

This competing relationship between bias and variance as model expressivity changes is known as the bias-variance tradeoff and understanding it is key to building powerful machine learning models. Usually an optimum is struck at an intermediate level of model complexity.

In practice, this optimum is determined by restricting a model's expressivity in a process known as regularization. This can be done directly by penalizing usage of more parameters. For example, a feedforward neural network with L2 regularization has a penalty on the norm of its parameters causing the model to only utilize its expressivity if it can fit the data better by some well defined degree. Other forms of regularization such as dropout, data augmentation, and adding noise to training data indirectly penalize overfitting. How strong regularization should be enforced is often validated by holding some data out at training and using it to validate the models performance on new data. Determining a model's regularization strength is one form of hyperparameter optimization that is computationally intensive and does not have a well-prescribed process.

1.2.2 Regularization via inductive biases

In addition to these regularization techniques, there are other ways to restrict a model's expressivity. The growing field of geometric deep learning restricts a model's expressivity in principled ways that stem from the physical nature of the dataset being learned [20]. In general, if physical symmetries are respected in the underlying nature of a system, then the model space can be restricted to a subspace of all possible models that respects these symmetries, with no concern that the true function is outside the model's expressivity. Said another way, respecting physical symmetries reduces the expressivity of the model and thus the variance but does not increase the bias. Thus one can expect such a model to have better performance on prediction since overfitting is less likely.

Models that incorporate physical symmetries are said to have inductive biases. Perhaps the most well-known case of inductive biases in neural networks is in convolutional neural networks (CNNs) in which images are processed by convolving a set of learned filters across the image. This convolutional process ensures that features such as edges are recognized equivalently no matter where they appear in an image. Coupled with max-pooling layers, these networks are ultimately symmetric to translations of an image.

Similarly, a neural network that operates on protein structure should be symmetric with respect to global rotations of atomic coordinates around a point of interest. A design

choice in building such a network involves how to incorporate this inductive bias into the network. One can make an extension of CNNs and use spherical filters that are convolved over 3D rotations; however in practical situations this only ensures symmetry up to discrete rotations. Furthermore, due to the different dimensionality, there are many more 3D rotations to perform in this convolution than in the two-dimensional translation case making this architecture much more expensive than in a traditional CNN. Alternatively, one can precompute rotationally symmetric quantities such as the total mass or pairwise displacements. However, this either restricts the model or embeds the 3D structure in an even higher-dimensional representation where the important correlations may not be as easy to uncover. A third option is to perform convolutions in frequency space where models can learn directly from the rich 3D data while symmetry is naturally accounted for if the correct operations are used. These models are elegant and efficient; however they do require some knowledge of group and representation theory. A great coverage of the topic is found in [26]. In the remaining section of this chapter, I will briefly provide background on these concepts to motivate our specific architecture choice and situate our model in the larger field of geometric deep learning.

1.3 Symmetry, Equivariance, and Geometric Deep Learning

1.3.1 Symmetry

Symmetry is ubiquitous in physics. The word symmetry in physics is usually used to refer to a transformation of a system which leaves the system exactly unchanged.[1] For instance, a square's symmetries are flips and rotations. These symmetries have well defined properties. Since one symmetry of a square is rotation by 90°, a square is consequently symmetric under all combinations of rotations by 90° (i.e., 180°, 270°, and 360°). This property along with others define a mathematical structure of a group. We will formalize symmetry groups in section 1.3.2.

Proteins are obviously more complicated than regular shapes like squares, and yet still,

[1] A system or quantity that is symmetric means that it has a symmetry and when we say that object is symmetric under a specific transformation, we mean that that transformation is a symmetry of the object.

quantities of interest still exhibit symmetries. For example, even though a protein generally is not a symmetric object, the energy of a conformation of a protein is symmetric under global rotations of the molecule. Analogously, even though a square is only symmetric under discrete rotations of 90°, the area of a square and the number of corners are symmetric under all continuous rotations. Ultimately, we will be interested in the energies of amino acids in atomic environments within proteins or at interfaces, and thus we will need symmetric features to approximate these energies.

Despite our priority of symmetry, we will need a larger understanding of how non-symmetric quantities behave under physical transformations to describe rich symmetric quantities. Consider the example of a square. The number of corners does not depend on the geometry of the square. However, the area does. Similarly, the number of atoms in a protein also does not depend on the geometry of the protein, but the energy does. This idea that symmetric quantities may depend on geometric features that are not necessarily symmetric themselves will be key to building a rich structure-to-function map. Symmetric quantities are called *invariant* and geometric quantities that are not symmetric under transformations but behave in some well-defined way are called *equivariant*.

The importance of equivariance has long been recognized in bottom-up models in physics and is recently being recognized in the development of top-down ML models. In fact, the growing field of geometric deep learning is the systematic study of how the concept of symmetry is used to build inductive biases into neural networks in order to reduce variance. To understand the methods emerging from this field and the method central to this thesis, I will first provide a minimal background on group theory and representation theory which will formalize the notions of physical transformations and how physical quantities change under transformations. This mathematical formalism will give us the tools necessary to build a robust geometric deep learning model that can learn from all-atom representations of protein structures.

1.3.2 Group and representation theory

To incorporate symmetry into our neural network in an efficient yet flexible way, we turn to the mathematics of representation theory which studies how groups act on vector spaces. First, we shall formalize the definition of a group for clarity.

Definition: A *group* $\mathcal{G}$ is a set of elements $g \in \mathcal{G}$ and a binary operation $\cdot : \mathcal{G} \times \mathcal{G} \to \mathcal{G}$ defined on the elements with four properties:

1. Closure: For all $g, h \in \mathcal{G}$, $g \cdot h \in \mathcal{G}$.
2. Identity: There exists $e \in \mathcal{G}$ such that $e \cdot g = g$ for all $g \in \mathcal{G}$.
3. Inverse: For all $g \in \mathcal{G}$, there exists a g^{-1} such that $g^{-1} \cdot g = e$.
4. Associativity: For all $g, h, k \in G$, $(g \cdot h) \cdot k = g \cdot (h \cdot k)$

It is easy to check that rotations satisfy these properties. For any rotation, a rotation about the same axis by the opposite angle is the inverse. For any two rotations applied to an object, there exists one rotation that maps the original pose to the final pose (closure). Finally associativity follows from the representation of rotations as 3×3 matrices and the associativity of matrix multiplication.

While a group is simply an abstract mathematical concept, we are concerned with how the outputs of functions (i.e., neural networks) transform under physical transformations of inputs. This brings us to the idea of a group representation.

Definition: A *representation* of a group G on a vector space V is a mapping $U : G \to \mathrm{GL}(V)$ that takes $g \in G$ to a linear operator $U(g)$ on a vector space V with the condition that the group multiplication rule is satisfied

$$U(g_1)U(g_2) = U(g_1 g_2). \tag{1.1}$$

Representations and the theory accompanying them are useful because they provide a description of a vector space on which a group acts. Our neural network will take as input the 3D structure of a protein and return a number. The input vector space of this function and ultimately the vector space we will be interested in is the vector space of signals over

3D space. Representation theory will provide us a natural basis for this vector space when we are prioritizing rotational equivariance.

Definition: A map $f : V_1 \to V_2$ is *equivariant* under a group $\mathcal{G}$ if for a representation U_1 on V_1 and U_2 on V_2, f satisfies the property

$$U_2(g)f(x) = f(U_1(g)x). \tag{1.2}$$

It turns out that the natural basis to use to to build group-equivariant maps over a vector space is provided by specific representations called the irreducible representations. We shall not completely cover irreducible representations here, but for a complete treatment suitable for one with a physics background see [138]. What is necessary to understand about irreducible representations is that they prescribe an orthonormal and complete basis over the vector space in consideration and therefore any operator U is block diagonalizable in this basis with each block corresponding to one irreducible representation.[2] If we use the basis provided by the irreducible representations, then by linearity of representations, we can linearly combine features belonging to the same representation and preserve equivariance.

ML models require nonlinear operations to build expressive models, but in general, nonlinear operations do not preserve equivariance. Luckily representation theory also provides a means of taking products and decomposing those product into equivariant parts. Ultimately, we will build a neural network that uses this equivariant basis along with symmetry respecting linear and nonlinear operations as precribed by representation theory. Since coverage of this theory requires more mathematics than is possible to cover here, we turn to two examples to illustrate the utility of irreducible representations.

Translational equivariance The irreducible representations of translations in one dimension are the complex exponentials e^{ikx}, and signals over the real line are famously written in this basis in the Fourier transform. Specifically, the Fourier transform of a signal on a real

[2]Physical observables are naturally classified according to irreducible representations. Consider scalar and vector quantities. While each transforms differently under rotations, a scalar remains a scalar and a vector remains a vector. These two types of quantities correspond to the two lowest dimensional representations of the rotation group. Quadrupole tensors correspond to the next.

line $f(x)$ can be expressed as $\hat{F}(k) = \int_{-\infty}^{\infty} e^{-ikx} f(x)dx$. Now, our rationale for using this basis is that it naturally classifies equivariant quantities under translations.

Consider a translated signal $f_a(x) = f(x-a)$ which is the original signal translated by a distance a. Clearly the new signal is generally not linear in the original signal at that same point. The Fourier transform of this translated signal follows

$$\begin{aligned}
\hat{F}_a(k) &= \int_{-\infty}^{\infty} e^{-ikx} f_a(x)dx \\
&= \int_{-\infty}^{\infty} e^{-ikx} f(x-a)dx \\
&= \int_{-\infty}^{\infty} e^{-ik(x'+a)} f(x')dx' \\
&= e^{-ika}\hat{F}(k).
\end{aligned}$$

Thus, the Fourier transform of the signal transforms as $\hat{F}(k) \xrightarrow{a} e^{-ika}\hat{F}(k)$, implying that the Fourier transform is equivariant under translation with an output operator that is simply a phase shift, i.e., $D_{\text{out}}(a) = e^{-ika}$. In three dimensions, translations act on vectors via $\mathbf{x} \to \mathbf{x} + \mathbf{a}$ and the representation is given by the operator $D_{\text{out}}(\mathbf{a}) = e^{-i\mathbf{k}\cdot\mathbf{a}}$. Thus, when considering spatial signals and prioritizing translational symmetry, the traditional Fourier basis of complex exponentials is a naturally equivariant basis. With this in mind, we can examine the natural equivariant basis when prioritizing rotational symmetry.

Rotational equivariance When considering rotations in 3D, we first note that the natural coordinate system to use is spherical coordinates since radii do not change under rotations. Thus, the domain of interest will be functions over the sphere $\chi(S^2)$. Similar to the translation case, we can use Fourier space to define an equivariant transformation for rotations. The basis for functions over the sphere defined by the irreducible representations is given by the spherical harmonics $Y_{lm}(\theta, \phi)$. Just as complex exponentials are sinusoidal functions over the real line, the spherical harmonics are sinusoidal functions over the sphere. The Fourier transform $\hat{F}$ of the signal on the sphere is given by

$$\hat{F}_{\ell m} = \int_0^{2\pi} \int_0^{\pi} f(\theta, \phi) Y_{\ell m}(\theta, \phi) \sin^2\theta \, \mathrm{d}\theta \, \mathrm{d}\phi \tag{1.3}$$

where $Y_{\ell m}(\theta, \phi)$ is the spherical harmonic of degree ℓ and order m defined as

$$Y_{\ell m}(\theta, \phi) = \sqrt{\frac{2\ell + 1}{4\pi} \frac{(\ell - m)!}{(\ell + m)!}} e^{im\phi} P_\ell^m(\cos\theta) \tag{1.4}$$

where ℓ is a nonnegative integer ($0 \leq \ell$), and m is an integer within the interval $-\ell \leq m \leq \ell$. $P_\ell^m(\cos\theta)$ is the Legendre polynomial of degree ℓ and order m, which, together with the complex exponential $e^{im\phi}$, defines sinusoidal functions over the angles θ and ϕ.[3]

Just as the Cartesian Fourier transform is well-behaved under translations, the spherical harmonics are similarly well-behaved under rotations. The operators that describe how spherical harmonics transform under rotations are called the Wigner D-matrices, denoted by $D^\ell_{mm'}(R)$ [138]; this operator is a matrix because the rotation group in 3D is not commutative. Under a rotation R, spherical harmonics transform as

$$Y_{\ell m}(\theta, \phi) \xrightarrow{R} \sum_{m'=-\ell}^{\ell} D^\ell_{m'm}(R) Y_{\ell m'}(\theta, \phi). \tag{1.5}$$

Since spherical harmonics transform in this well-defined way, we can now examine how a rotation acts on the Fourier transform of a signal over the sphere. Suppose we have a signal $f(\mathbf{n})$ defined on the elements $\mathbf{n}$ of the spherical shell S^2 (i.e., the set of angular coordinates for points on the sphere). The Fourier coefficients associated with the rotated signal $f_R(\mathbf{n}) = f(R\mathbf{n})$ follow,

$$\begin{aligned}
\hat{F}^R_{\ell m} = \int Y_{\ell m}(\mathbf{n}) f_R(\mathbf{n}) \mathrm{d}\Omega &= \int Y_{\ell m}(\mathbf{n}) f(R\mathbf{n}) \mathrm{d}\Omega \\
&= \int Y_{\ell m}(R^{-1}\mathbf{n}') f(\mathbf{n}') \mathrm{d}\Omega' \\
&= \int \sum_{m'=-\ell}^{\ell} D^\ell_{m'm}(R^{-1}) Y_{\ell m}(\mathbf{n}') f(\mathbf{n}') \mathrm{d}\Omega' \\
&= \sum_{m'=-\ell}^{\ell} D^\ell_{m'm}(R^{-1}) \int Y_{\ell m}(\mathbf{n}') f(\mathbf{n}') \mathrm{d}\Omega' \\
&= \sum_{m'=-\ell}^{\ell} D^\ell_{m'm}(R^{-1}) \hat{F}_{lm}
\end{aligned}$$

[3] In quantum mechanics, spherical harmonics are used to represent the orbital angular momenta, e.g., for an electron in a hydrogen atom. In this context, the degree ℓ relates to the eigenvalue of the square of the angular momentum, and the order m is the eigenvalue of the angular momentum about the azimuthal axis.

where $\mathrm{d}\Omega = \sin^2\theta\,\mathrm{d}\theta\,\mathrm{d}\phi$ is the angular differential in the spherical coordinate system. Under a rotation R about the origin, the spherical Fourier coefficients of a real-space signal transform according to $\hat{F}_{\ell m} \xrightarrow{R} \sum_{m'=-\ell}^{\ell} D^{\ell}_{m'm}(R^{-1})\hat{F}_{\ell m'}$, which is simply a matrix product. Thus, the spherical harmonics provide a natural equivariant basis to use for describing signals over the sphere.

Nonlinear transforms respecting rotational equivariance One key feature of neural networks is applying nonlinear activations, which enable a network to approximately model complex nonlinear phenomena. Commonly used nonlinearities include reLU, tanh, and softmax functions. However, these conventional nonlinearities can break rotational equivariance. To construct expressive rotationally equivariant neural networks we can use the Clebsch-Gordan tensor product, which is the natural nonlinear (or bilinear in the case of using two sets of Fourier coefficients) operation in the space of spherical harmonics [138].

Given two spherical tensors $\hat{F}_{\ell_1}$ and $\hat{G}_{\ell_2}$, we can take a product between all of their components $\hat{F}_{\ell_1 m_1}\hat{G}_{\ell_2 m_2}$, which would allow us to express nonlinearities. Although these products do not behave well under rotations, linear combinations of them do transform in well-defined ways. The Clebsch-Gordan coefficients are the coefficients that decompose these products back into the space of spherical harmonics, where Wigner D-matrices define the rotation operations. Specifically, the product of spherical tensors $\hat{F}_{\ell_1}$ and $\hat{G}_{\ell_2}$ will yield spherical tensors of degree L such that $|\ell_1 - \ell_2| < L < \ell_1 + \ell_2$ in the following way:

$$\hat{H}_{LM} = \sum_{m_1=-\ell_1}^{\ell_1} \sum_{m_2=-\ell_2}^{\ell_2} \langle LM|\ell_1 m_1; \ell_2 m_2\rangle \hat{F}_{\ell_1 m_1}\hat{G}_{\ell_2 m_2} \tag{1.6}$$

where $\langle LM|\ell_1 m_1; \ell_2 m_2\rangle$ is a Clebsch-Gordan coefficient and can be precomputed for all degrees of spherical tensors [138]. Similar to spherical harmonics, Clebsch-Gordan products also appear in quantum mechanics, and they are used to express couplings between angular momenta. In following with recent work on group-equivariant machine learning [70, 135, 10, 93], we will use Clebsch-Gordan products to express nonlinearities in 3D rotationally equivariant neural networks for protein structures.[4]

[4]These are two examples of how irreducible representations can be used to make a convolutional neural

The upshot of this illustrative example is that if we can encode atomic point clouds in the rotationally equivariant spherical harmonic basis, then we can build equivariant neural networks from linear and nonlinear transformations detailed above. Thus, we have our mathematical foundation from which we will attempt to build a geometric deep learning model of protein function from structure. Before presenting the rest of this thesis, let me summarize the entirety of this chapter.

1.4 Summary

First, we established that proteins utilize a small library of amino acids to accomplish a diverse number of functions. It is the unique composition of these amino acids' side chains in 3D space and the potential for interactions with the molecular environment that gives rise to any protein's function. Understanding the interactions between these amino acids and their atomic environments is key to understanding how protein function arises from structure. Computational prediction of how well proteins perform their function based on their structure is important but difficult due to the high dimensionality of 3D structures. In this age of growing computational power, machine learning offers attractive tools to build robust data-driven models, but generally lacks awareness of symmetry that allows for the interpretability and generalizability of physical models. The field of geometric deep learning prescribes how to build a symmetry-aware model using the math of representation theory, with one option being a fully Fourier space treatment of signals. In the next chapter, I will follow this prescription in creating the model central to this thesis: the Holographic Convolutional Neural Network.

---

network. In fact, a more general theory of group-convolutional neural networks has been developed. While it is outside the scope of this thesis, an interested reader should explore [26]

A brief note on publications

My Ph.D. work has resulted in three manuscripts: one that is directly contained in this thesis (specifically chapters 2 and 3) and two that are not. For completeness, here are the three works:

- Pun, M. N., Ivanov, A., Bellamy, Q., Montague, Z., LaMont, C., Bradley, P., Otwinowski, J., & Nourmohammad, A. (2022). *Learning the shape of protein micro-environments with a holographic convolutional neural network* (p. 2022.10.31.514614). bioRxiv. https://doi.org/10.1101/2022.10.31.514614

- Visani, G. M., Pun, M. N., & Nourmohammad, A. (2022). *Holographic-(V)AE: An end-to-end SO(3)-Equivariant (Variational) Autoencoder in Fourier Space* (arXiv:2209.15567). arXiv. https://doi.org/10.48550/arXiv.2209.15567

- Nourmohammad, A., Pun, M., & Visani, G. M. (2022). *Machine-Learning Model Reveals Protein-Folding Physics.* Physics, 15, 183. https://doi.org/10.1103/PhysRevLett.129.238101

Chapter 2

HOLOGRAPHIC CONVOLUTIONAL NEURAL NETWORK

This chapter contains work from the following paper which is in submission at the time of writing:

> Pun, M. N., Ivanov, A., Bellamy, Q., Montague, Z., LaMont, C., Bradley, P., Otwinowski, J., & Nourmohammad, A. (2022). *Learning the shape of protein micro-environments with a holographic convolutional neural network* (p. 2022.10.31.514614). bioRxiv. https://doi.org/10.1101/2022.10.31.514614

In the previous chapter, I detailed how a protein's function is dictated by physical interactions in 3D space between amino acid side chains and the surrounding atomic environment. I noted how prediction of quantitative function from structure is a challenging problem and the field of geometric deep learning offers a new set of tools to build attractive models for predicting a protein's quantitative function from its structure. Specifically these equivariant models can learn rich features from geometric data and promise to build more efficient, interpretable, and generalizable data-driven models.

In this chapter, I will detail the neural network central to this thesis—the Holographic Convolutional Network (H-CNN). I will develop this network to solve the task of predicting an hidden amino acid's identity based on the atomic surroundings. This task is formulated in an attempt to learn amino acid propensities in atomic environments and later use these learned amino acid propensities as a modular building block of a larger structure-to-function map. First, in section 2.1, I will formulate the neural network focusing on the rotationally equivariant encoding of atomic point clouds and the network's rotationally equivariant architecture. Then, in section 2.2, I will summarize the model's performance and how the general structure of its predictions reflect known chemistry and evolutionary use of amino acids.

2.1 Rotationally equivariant structure-to-function map for proteins

2.1.1 Rotationally equivariant encoding of protein micro-environments as spherical holograms

We define an amino acid's micro-environment as all of the atoms associated with the surrounding residues within a 10Å of the central residue's alpha carbon. We use the position of the central residue's alpha carbon as the reference point of the neighborhood $\mathcal{N}$. Each atom i within this neighborhood, located at position $\mathbf{r}_i$ with respect to the reference point, has three attributes associated with it: (i) the element type of the atom, $e_i \in \{C, N, O, S, H\}$, (ii) the partial charge of the atom $Q_i \in \mathbb{R}$, and (iii) the SASA of the atom $A_i \in \mathbb{R}_0^+$. With these characteristics we define a feature vector v_i^c where c indexes the input channels $\{C, N, O, S, H, \text{charge}, \text{SASA}\}$. This vector takes on values given by

$$v_i^c = \begin{cases} \delta_{e_i,c} & \text{for } c \in \{C, N, O, S, H\} \\ Q_i & \text{for } c = \text{charge} \\ A_i & \text{for } c = \text{SASA} \end{cases}$$

where δ_{ij} is a Kronecker delta function which takes the value 1 if $i = j$ and 0 otherwise. Thus, for the elemental channels (i.e., $c \in \{C, N, O, S, H\}$) the feature vector v_i^c takes value 1 if atom i is of type c and 0 otherwise. Q_i and A_i denote the partial charge and the SASA (in units of Å^2) of atom i, respectively. These quantities are computed by PyRosetta for each protein structure [22].

Each channel defines a point cloud with the attributed values stored at the coordinates of the corresponding atoms. We express the point cloud of each channel in a spherical coordinate system with the origin set at the alpha carbon of the central amino acid in the neighborhood. We then define an atomic density as a sum of Dirac delta distributions parameterized by each atomic coordinate in the neighborhood $\rho^c(\mathbf{r}) = \sum_{i \in \mathcal{N}} v_i^c \delta^{(3)}(\mathbf{r} - \mathbf{r}_i)$. Here $\delta^{(3)}(\mathbf{x})$ denotes a normalized Dirac delta probability density in 3D, which is zero everywhere except at the origin $\mathbf{x} = \mathbf{0}$, where it is infinite. We use 3D Zernike transforms to encode the point clouds associated with each channel in a 3D rotationally equivariant

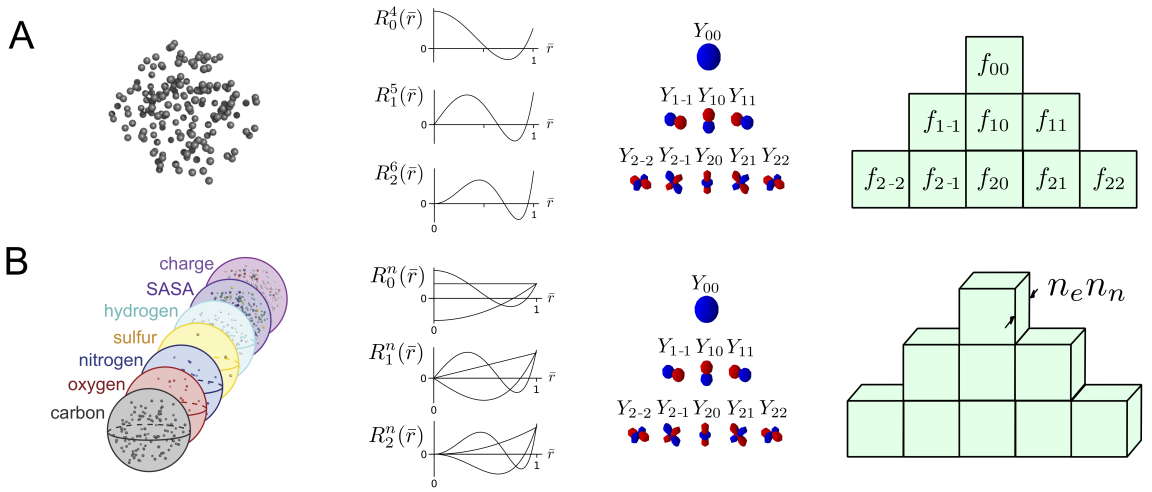


Figure 2.1: **Rotationally equivariant encoding of atomic neighborhoods with the 3D Zernike transform. (A)** Any point cloud (left) can be written in the equivariant spherical harmonic Fourier basis by integrating the point cloud density against the spherical harmonics along with a choice of a radial function (middle), which depends on the spherical harmonic degree ℓ. The triangular structure (right) demonstrates the resulting Fourier coefficients $f_{\ell m}$, in which the rows reflect the different degrees of spherical harmonics $0 \leq \ell$, which are nonnegative integers. Individual cells within each row correspond to different orders m of the spherical harmonics, which are integers bounded by $-\ell \leq m \leq \ell$. **(B)** The Fourier transform becomes three-dimensional when radial resolution is required thereby increasing the number of discrete n-values used n_n. The multiplicity of coefficients is also determined by the number of point clouds being transformed n_ρ (e.g.,the different atomic channels on the left). Since both the point clouds and the radial information are invariant under rotations, these two dimensions can be combined in index to give preference in organizing information according to the degree ℓ and order m of the spherical harmonics.

fashion in the spherical Fourier space:

$$\hat{Z}^c_{n\ell m} = \int \rho^c(\mathbf{r}) Y_{\ell m}(\theta, \phi) R_{n\ell}(\mathbf{r})\, \mathrm{d}\Omega. \tag{2.1}$$

In equation 2.1, $\mathrm{d}\Omega$ is the angular differential in the spherical coordinate system, $Y_{\ell m}(\theta, \phi)$ is the spherical harmonics (eq. 1.4), and $R_{n\ell}(\mathbf{r})$ is the radial function of the 3D Zernike transform, which can be expressed as

$$R_{n\ell}(\mathbf{r}) = (-1)^{\frac{n-\ell}{2}} \sqrt{2n+3} \binom{\frac{n+\ell+3}{2} - 1}{\frac{n-\ell}{2}} |\mathbf{r}|^\ell \, {}_2F_1\left(-\frac{n-1}{2}, \frac{n+\ell+3}{2}; \ell + \frac{3}{2}; |\mathbf{r}|^2\right) \tag{2.2}$$

where ${}_2F_1(\cdot)$ is the ordinary hypergeometric function. The index $n \geq 0$ is a nonnegative integer, and the radial function $R_{n\ell}(\mathbf{r})$ is nonzero only for even values of $n - \ell \geq 0$. Zernike polynomials form a complete orthonormal basis in 3D and, therefore, can be used to expand and retrieve any 3D shape if large enough ℓ and n coefficients are used. Approximations that restrict the series to finite n and ℓ are often sufficient for shape retrieval, and hence, desirable algorithmically. Importantly, expansion of an input signal by Zernike polynomials (eq. 2.1) is rotationally equivariant since a rotation R transforms the resulting coefficients through Wigner-D matrices as $\hat{Z}^c_{n\ell m} \xrightarrow{R} \sum_{m'=-\ell}^{\ell} D^\ell_{m'm}(R^{-1}) \hat{Z}^c_{n\ell m'}$ (eq. 1.5).

Zernike projections in the spherical Fourier space can be thought of as a superposition of spherical holograms of an input point cloud, and thus, we term this operation the *holographic encoding* of protein micro-environments.

2.1.2 Rotationally equivariant model of protein micro-environments with H-CNN

We use the coefficients of the Zernike expansion $\hat{Z}^c_{nlm}$ in eq. 2.1 (i.e., the holographic encoding of the data) as inputs to a rotationally equivariant convolutional neural network to classify amino acids based on their surrounding atomic micro-environments. We term this network holographic convolutional neural network (H-CNN).

The convolutional neural network that we use for our analysis is a Clebsch-Gordan network (CGNet), described in ref. [70]. CGNet is built out of three rotationally equivariant modular units: (i) linear transformation of spherical harmonics, (ii) Clebsch-Gordan nonlinearity, and (iii) spherical batch normalization. Below we will describe the computations done in each of these modular equivariant units. Without loss of generality, we denote the operations in

each unit of the network in terms of a general equivariant signal $\hat{F}^k_{lm}$ where k is a catch-all index that represents all nonrotational indices (i.e., channel c and index n in eq. 2.1), and ℓ and m correspond to the angular indices (eq. 1.4). The output of each unit will be denoted by $\hat{G}^k_{lm}$.

Linear transformations in CG-nets. Linearities are ubiquitous in neural networks and contribute partially to their powerful expressiveness. Our linearity takes the form

$$\hat{G}^k_{\ell m} = \sum_{k'} W^{(\ell)}_{kk'} \hat{F}^{k'}_{\ell m} \tag{2.3}$$

where $W^{(\ell)}_{kk'}$ denotes two-dimensional matrices that mix different input channels k' (e.g., combinations of radial index n and channel c in the input or general learned representations in intermediate layers) producing different output channels k. To preserve rotational symmetry, we will impose the restriction of only combining information (i.e., forming linear combinations of inputs) that belongs to the same irreducible representations (irreps) of SO(3) in eq. 2.3. In other words, we mix only inputs associated with the same degree ℓ and order m of spherical harmonics. This constraint, coupled with the linearity of the operators in the definition of equivariance (i.e., the Wigner D-matrices), guarantees that the outputs of a linearity are rotationally equivariant as well.

Clebsch-Gordan nonlinearity. As mentioned in section 1.2, nonlinearities are ubiquitous in neural networks and are key to their expressivity. The Clebsch-Gordan nonlinearity is the simplest rotationally equivariant nonlinearity, and it is expressive enough for most learning tasks [70, 93]. This Clebsch-Gordan nonlinearity is simply a product of spherical tensors that is decomposed back into the spherical harmonic basis via the Clebsch-Gordan coefficients,

$$\hat{G}^K_{LM} = \sum_{m_1=-\ell_1}^{\ell_1} \sum_{m_2=-\ell_2}^{\ell_2} \langle LM|\ell_1 m_1; \ell_2 m_2\rangle \hat{F}^{k_1}_{\ell_1 m_1} \hat{F}^{k_2}_{\ell_2 m_2} \tag{2.4}$$

where $\langle LM|\ell_1 m_1; \ell_2 m_2\rangle$ is a Clebsch-Gordan coefficient and $K = (k_1, k_2)$ is the new channel index. We echo Kondor's original remark in [70] that although the Clebsch-Gordan product is not as nonlinear as many other nonlinearities used in state-of-the-art machine learning

models and thus may appear to be less expressive, we find it to be sufficiently expressive in CGNets for all of our purposes in this thesis.

Generally we have a choice for which combinations of values k_1, k_2, ℓ_1, and ℓ_2 are used in this product. We focus on two choices for these indices. We define *fully connected* networks, for which k_1, k_2, ℓ_1, and ℓ_2 are allowed to take on all possible values. We also define *simply connected* networks, for which we impose the condition that $k_1 = k_2$ and $\ell_1 = \ell_2$. The exact choice made in combining these indices affects the dimensionality of the network. The width of the Clebsch-Gordan nonlinearity output in a fully connected network with a maximum spherical degree $L_{\max}$ and a width d_h (i.e., $k \in \{1, ..., d_h\}$) is given by

$$\Omega^{\text{full}}_{\ell,L_{\max}} = d_h^2\left(\left[\frac{1}{4}(2+\ell)(5+\ell) - 2\left\lfloor\frac{1}{2}(\ell-1)\right\rfloor\right] + (\ell+1)(L_{\max}-\ell)\right) + (\ell+1)(L_{\max}-\ell). \quad (2.5)$$

Meanwhile, for a simply connected network with the same $L_{\max}$ and d_h, the width of the Clebsch-Gordan nonlinearity is

$$\Omega^{\text{simple}}_{\ell,L_{\max}} = d_h(L_{\max} + 1 - \ell). \quad (2.6)$$

Spherical batch normalization. Batch normalization is a common feature in neural networks that allows for smooth training of the network. Since the output of our nonlinear Clebsch-Gordan product is not bounded, batch normalization becomes extremely important to ensure that activations remain finite throughout training. We impose a batch normalization layer after each linear operation, producing normalized inputs to the nonlinear operation (Clebsch-Gordan product). We use the following batch normalization during training:

$$\hat{G}^k_{\ell m} = \frac{\hat{F}^k_{\ell m}}{\sqrt{\langle|\hat{F}^k_\ell|^2\rangle_{\text{batch}} + \epsilon}} \quad (2.7)$$

where $|\hat{F}^k_\ell|^2 = \sum_{m=-\ell}^{\ell} \hat{F}^k_{\ell m}\hat{F}^{*k}_{\ell m}/(2\ell+1)$, * denotes complex conjugation, $\langle\cdot\rangle_{\text{batch}}$ denotes averaging over a batch, and ϵ is a small number used to ensure numerical stability, which we set to $\epsilon = 10^{-3}$. During training, the network stores a moving average of the norm for each ℓ in each channel as

$$N^{k,i}_\ell = \xi\langle|\hat{F}^k_\ell|^2\rangle_{\text{batch}} + (1-\xi)N^{k,i-1}_\ell \quad (2.8)$$

where ξ serves as the momentum of our moving average, set to $\xi = 0.99$. This moving average is then used as the norm during testing in the following way:

$$\hat{G}^k_{\ell m} = \frac{\hat{F}^k_{\ell m}}{\sqrt{N^{k,i}_\ell + \epsilon}} \tag{2.9}$$

The different batch normalization used for training and testing prevents the batching of the validation data to affect the evaluation of the model accuracy during testing.

CG layer A Clebsch-Gordan (CG) layer is comprised of one linear operation, one spherical batch normalization, and one nonlinear CG product in that order. Computationally, a concatenation is also performed to collect all outputs of the CG product into a unified set of spherical Fourier coefficients. A schematic of a CG layer is provided in Fig. 2.2A. A full H-CNN then consists of multiple CG layers in series with each other. After the input Fourier representation has been passed through all layers, invariant information is collected from each layer (pre-sorted in the $\ell = 0$ components thanks to our prioritization of symmetry in using the irrep basis). These invariants are then broken into real and imaginary parts and fed into a series of dense linear layers which project the data into a 20-dimensional space. Finally, this 20-dimensional vector parameterizes a Boltzmann distribution over the 20 amino acids which define the H-CNN's predictions for each amino acid appearing in the center of the neighborhood. In the language of ML, this operation is equivalent to applying a softmax to the 20-dimensional output. We emphasize that besides this softmax, the only nonlinearity used in the network is the CG product meaning that the network's predictions are scalar polynomials of degree $n+1$ in the input Fourier coefficients where n is the number of layers.

2.2 Model performance

2.2.1 Amino acid classification

H-CNN was trained on neighborhoods from all evolutionarily unrelated crystal structures in the Protein Data Bank (PDB) [12]. The model was trained to minimize the categorical cross-entropy loss between the predicted probabilities and a one-hot distribution reflecting

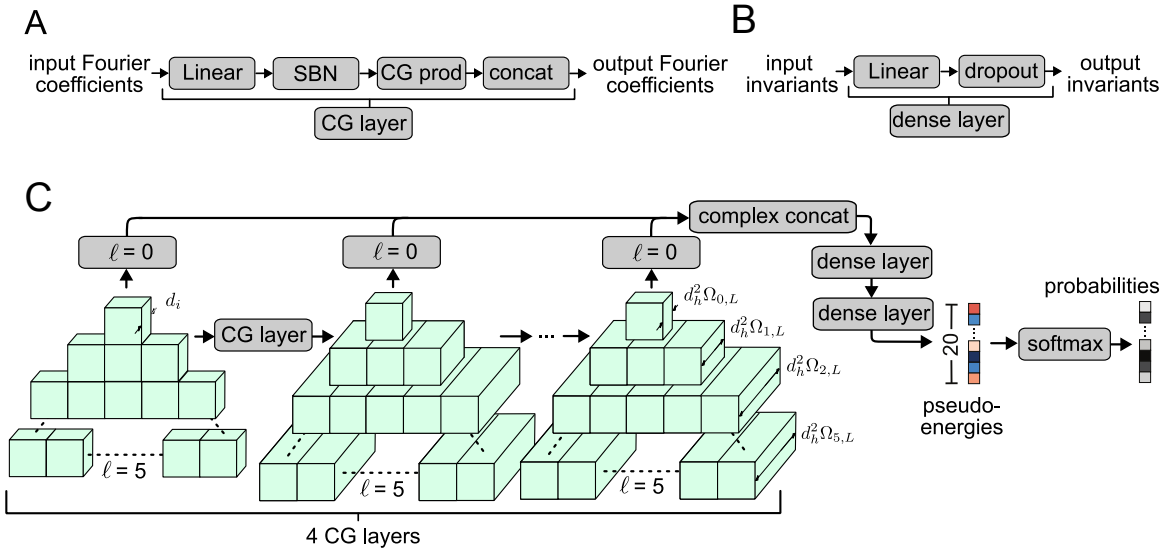


Figure 2.2: **Schematic of full H-CNN model architecture. (A)** The Clebsch-Gordan layer (CG layer) is fully equivariant and is comprised of a linear layer, a spherical batch norm (SBN), Clebsch-Gordan product (CG prod), and a degree-wise concatenation (concat). **(B)** Throughout all layers of the network, all invariants ($\ell = 0$) are collected and fed into a series of dense layers which are simple linear combinations with training dropout. **(C)** These two main components in (A,B) comprise the bulk of the entire network. First the input Fourier coefficients are processed through successive CG layers. Invariants are collected from the original input and from the output of every CG layer. The real and imaginary components of these invariants are split and concatenated (complex concat) and then fed into a series of dense layers. This schematic reflects the dimensions of the optimal model discovered in this work. Four CG layers were used. We note that the maximal width of the network as determined by the output of the nonlinearity depends on ℓ due to the selection rules of the CG prod. We denote this width $d_h^2\Omega_{\ell,L}$ where L is the maximal degree ℓ used in the network and are given in equations 2.5 and 2.6. Two dense layers were used in the optimal model. The dimensions of these layers are (4852×500) and (500×20). The output of these dense layers are termed pseudo-energies and act as logits in a softmax to estimate probabilities for each amino acid. We emphasize the only nonlinearities in the network are the Clebsch-Gordan products and the ultimate softmax.

the true central amino acid. For more information on the dataset, training procedure, and hyperparameter optimization, see Appendix A. For both the simply connected and the fully connected networks, we found the best hyperparameters that minimize the validation cross-entropy loss function (see equation A.1). The best simply connected model had 62% classification accuracy and a minimal validation loss of 1.2 while the best fully-connected model had 68% classification accuracy and a minimal validation loss of 0.91.[1] This difference in predictive power of the models appeared consistent across all training scenarios.

We also note that the best H-CNN needed little regularization beyond its inductive biases with regularization strength of $\lambda_r = 1.2 \times 10^{-16}$ and dropout rate $\lambda_{dr} = 5.49 \times 10^{-4}$ (Table A.1). This supports the hypothesis that the inductive biases built into this minimal symmetry-aware model coupled with just early stopping of training are sufficient for preventing overfitting.

We compare the accuracy and the training efficiency of our model to existing methods that classify amino acids based on all-atom representations of their local environments in Table 2.1. For this comparison, we used the metric of testing accuracy of the amino acid class since this was more commonly reported in the literature. Specifically, we compare H-CNN with two 3D CNN methods that used conventional translationally convolutional neural networks on voxelized protein structures [137, 117]. Notably, these 3D CNN methods require local alignment of neighborhoods to the central amino acid. Also, each of these methods uses a cube of side length 20Å, while spheres of radius 10Å are used as inputs to H-CNN. Our method thus sees approximately half ($\pi/6$) of the volume that these other models see. Overall, our model has a comparable accuracy to the best 3D CNN method [117], but it is much more efficient in training.

We also compare our model to other methods that prioritize rotational symmetry, i.e., the Spherical CNN introduced in ref. [17] and the 3D Steerable CNN from ref. [150]. Spherical CNN [17] is approximately rotationally invariant by performing 3D convolutions on a discretized grid over the 3D sphere. 3D Steerable CNN [150] is a rotationally invariant method by applying spherical convolutions in a steerable basis. H-CNN performs better than

[1] An untrained model is expected to have a loss of $\log 20 \approx 3$.

method	rotationally invariant	dataset	post-processing	N	No. of parameters	training time	accuracy
H-CNN	yes	ProteinNet	charge hydrogen SASA	2.8×10^6	3.6×10^6	4.54 hours	68%
3D CNN [137]	no	SCOP & ASTRAL	None	7.2×10^5	10^7	3 days	40%
3D CNN [117]	no	SCOP & ASTRAL	PDB-REDO[65] charge hydrogen SASA	1.6×10^6	6.1×10^7	–	70%
Spherical CNN [17]	approx.	PISCES	charge hydrogen	–	6×10^7	–	56%
Steerable CNN [150]	yes	SCOP & ASTRAL	PDB-REDO charge hydrogen SASA	1.6×10^6	3.3×10^7	–	58%

Table 2.1: **Comparison of structure-informed models for protein neighborhoods.** H-CNN and existing methods trained on classifying all-atom neighborhoods are listed along with the summary quantities of the models. H-CNN demonstrates the power of respecting symmetry since it has fewer parameters and trains faster than 3D-CNNs despite being trained on at least the same amount of data.

both of these methods in the classification task with an order of magnitude fewer parameters (Table 2.1).

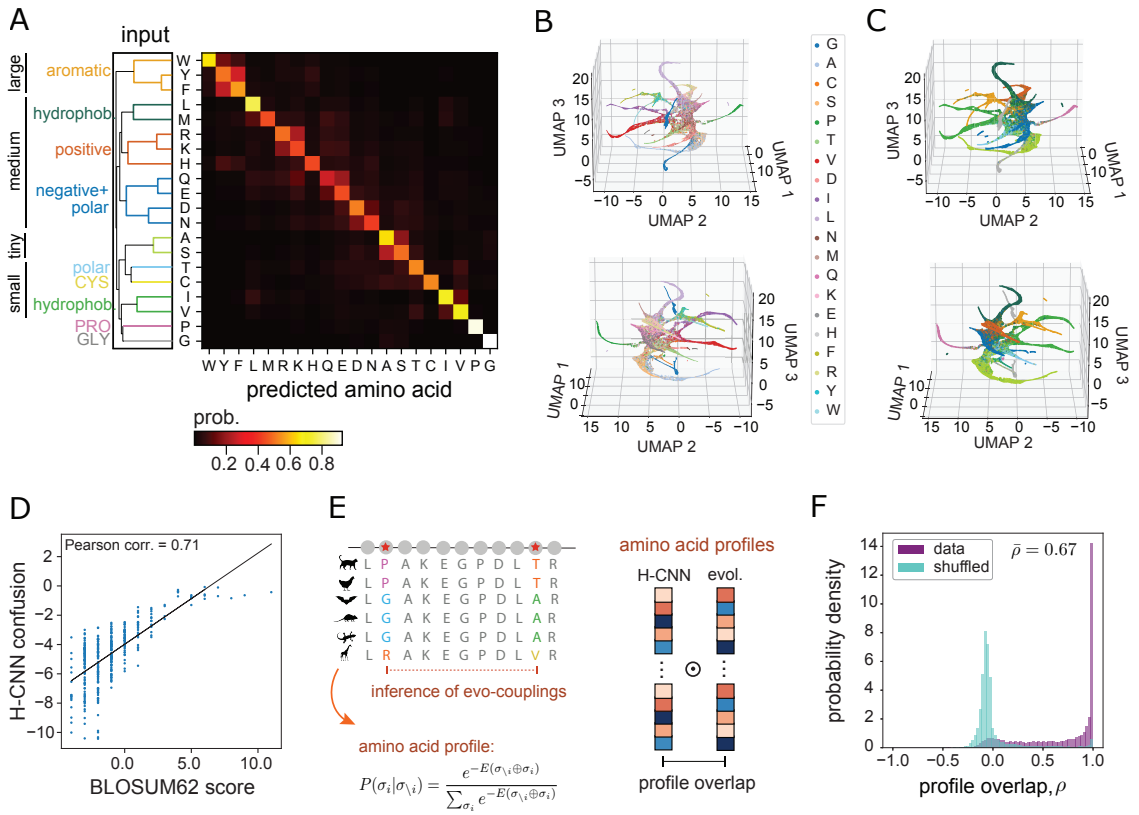
2.2.2 H-CNN reveals physicochemical properties of amino acids, consistent with evolutionary variation

We evaluated H-CNN's behavior beyond the accuracy of classification by analyzing the structure of H-CNN's predictions. H-CNN predicts the conformationally unique amino acids of glycine and proline with over 90% accuracy. Meanwhile, amino acids with typical side chains cluster based on their sizes and the physicochemical properties of the side chains including aromatic, hydrophobic, and charged groupings (Fig. 2.3A). The inferred amino acid preferences cluster well according to the input amino acid type (true label) in the low-dimensional UMAP representation [86], and amino acids with similar physicochemical properties cluster in nearby regions in the UMAP (Fig. 2.3B,C).

The reflection of physicochemical properties in H-CNN's predictions is further evidenced in ablation studies. We studied the effect of our post-processing of atomic neighborhoods by performing ablation studies on charge and SASA. For each ablation, we reran the second stage of our hyperparameter optimization and only considered fully connected networks. The overall accuracy from removing SASA was 57% while the overall accuracy from removing charge was 56%, in contrast to the 68% accuracy of the complete model shown in Fig. 2.3. A similar 10% drop in accuracy associated with SASA and charge was previously reported in ref. [117]. Fig. 2.4 shows the confusion matrices for each model compared to the best model's confusion matrix.

The results of these ablation studies further reveal that H-CNN's processing of information corresponds to physical intuition. Removing information about SASA or charge from the input data results in roughly a 10% drop in classification accuracy. Information from SASA mostly impacts the network's ability to predict hydrophobic amino acids, with some hydrophilic amino acids (R,K,E) also impacted. When charge is removed, the network demonstrates worse predictions on charged and polar amino acids most notably R, C, N, and E.

Figure 2.3: **H-CNN predicts amino acid preferences in protein microenvironments.** **(A)** The confusion matrix for amino acid predictions with H-CNN shows the mean H-CNN predicted probabilities of each of the twenty amino acids (output) conditioned on a specific central amino acid (input). Overall prediction accuracy is 68%. The hierarchical clustering for these predictions reflects known similarities in size and physicochemical properties of amino acids. **(B,C)** Low-dimensional projections of the prediction outputs (3D UMAPs) are shown. UMAPs are annotated by **(B)** the amino acid types, and **(C)** the physicochemical clusters in (A), with panels showing a different view of the UMAP in each case. Neighborhoods are closely clustered by amino acid types (B), and are spatially arranged based on the physicochemical properties of the side chains (C); colors in (C) are consistent with (A). **(D)** Amino acid confusion in (A) correlates with the substitutability of amino acids in natural proteins as determined by the BLOSUM62 matrix; 71% Pearson correlation. **(E)** Schematic shows how evolutionary covariation of amino acids in multiple sequence alignments of protein families can be used to fit Potts models (EV-couplings [56]) to characterize the probability of a given amino acid, given the rest of the sequence (left); see section 2.2.3 for details. To compare evolutionary and H-CNN predictions for site-specific amino acid profiles, the profile overlap is computed as the centered cosine similarity between the predicted probability profiles (right); see equation 2.14. **(F)** The profile overlaps are strongly peaked around one, implying perfect overlap in data (purple); the average profile overlap across 11,221 sites from a total of 67 protein families is $\bar{\rho} = 0.67$. H-CNN's predictions are notably different for the shuffled data, for which the profile overlap peaks near zero (cyan), with an average of 0.002.


A
input
large
medium
tiny
small
aromatic
hydrophob.
positive
negative+ polar
polar
CYS
hydrophob.
PRO
GLY
W Y F L M R K H Q E D N A S T C I V P G
predicted amino acid
prob.
0.2 0.4 0.6 0.8
B
C
UMAP 1
UMAP 2
UMAP 3
G A C S P T V D I L N M Q K E H F R Y W
D
Pearson corr. = 0.71
H-CNN confusion
BLOSUM62 score
E
inference of evo-couplings
amino acid profile:
$P(\sigma_i|\sigma_{\setminus i}) = \frac{e^{-E(\sigma_{\setminus i} \oplus \sigma_i)}}{\sum_{\sigma_i} e^{-E(\sigma_{\setminus i} \oplus \sigma_i)}}$
amino acid profiles
H-CNN
evol.
profile overlap
F
data
shuffled
$\bar{\rho} = 0.67$
probability density
profile overlap, ρ

2.2.3 H-CNN predictions reflect evolutionary usage of amino acids

H-CNN predictions reflect amino acid preferences seen in evolutionary data, even though the network is not trained on multiple sequence alignments (MSAs) of protein homologs. Specifically, the interchangeability of amino acids that H-CNN predicts is 71% correlated with the substitution patterns in evolutionary data, represented by the BLOSUM62 matrix (Fig. 2.3D). This correlation means that the average confusion of H-CNN reflects the true substitutability of amino acids when averaged over sites.

To further determine the alignment of H-CNN predictions with evolutionary use of amino acids on a site-by-site basis, we compared the predictions to the substitutions found in multiple sequence alignments (MSAs). We first noted that the substitutions in MSAs are not necessarily single point mutations. In general, substitutions at multiple sites can occur simultaneously as well as alongside insertions, deletions, and rearrangements for different isoforms depending on the sequence similarity threshold used in the MSA. If we consider the frequency at which amino acid α occurs at site i of an MSA, we see that such a frequency implicitly marginalizes over all other possible sequences. Specifically the probability of observing amino acid α at site i can be written as

$$P_i(\alpha) = \sum_{\sigma_{/i}} P_i(\alpha|\sigma_{/i})P(\sigma_{/i}) \tag{2.10}$$

where $\sigma_{/i}$ represents a sequence of amino acids at all sites but i and is summed over all possible sequences.

This probability is markedly different than the probability that H-CNN produces $P_\theta(\alpha|\mathbf{x})$ where the conditional variable $\mathbf{x}$ assumes a fixed sequence over residues that neighbor site i in 3D space. In genetics terminology, H-CNN does not account for epistasis whereby a mutation at one site may allow a previously unfavorable amino acid to be plausible in a sequence. Since H-CNN is limited in this way, we attempted to account for epistatic effects in MSA by inferring a Potts model from MSAs. A Potts model is a generalization of an Ising model where the state space of each node in the system has more than two possible states. In the case of proteins, the state space is 20-dimensional with the states being the 20 amino acids. Similar to the Ising model, the energy of a state is assumed to have two terms:

one due to the coupling of the state with some background field and another accounting for the couplings of sites with each other. Specifically for a sequence σ, the energy is written as

$$E(\sigma) = -\sum_i h_i(\sigma_i) - \sum_{i,j<i} J_{ij}(\sigma_i, \sigma_j). \tag{2.11}$$

The probability of a sequence σ is then simply given by the Boltzmann distribution

$$P(\sigma) = \frac{e^{-E(\sigma)}}{\sum_{\sigma'} e^{-E(\sigma')}}. \tag{2.12}$$

A maximum entropy approach can be used to infer the field and coupling terms and has been shown to both reflect structural interactions and account for general higher-order correlations in protein structures [149, 91]. After inference of a model, we can directly compare H-CNN probabilities $P_\theta(\alpha|\mathbf{x})$ with Potts model probabilities

$$P(\alpha|\sigma_{/i}) = \frac{e^{-E(\sigma_1\sigma_2\ldots\sigma_{i-1}\alpha\sigma_{i+1}\ldots\sigma_L)}}{\sum_\beta e^{-E(\sigma_1\sigma_2\ldots\sigma_{i-1}\beta\sigma_{i+1}\ldots\sigma_L)}} \tag{2.13}$$

where α is summed over the 20 possible amino acids.

We used EV-couplings [56] to infer Potts models for each of the domains in our testing set. For each protein chain in our test data, we used EV-Couplings to compute the MSA for the protein associated with the UniProt ID listed under the associated PDB entry. The MSA was created using UniRef90 [134] that clusters homologous proteins such that sequences within each cluster have at least 90% identity to and 80% overlap with the longest sequence (seed). For inference of the Potts model, we used the default EV-Couplings settings for the rest of the parameters. Ultimately we were able to infer evolutionary models on 67 protein families corresponding to 67 chains in our testing set. These protein families amounted to 11,221 unique sites, for which we compare the EV-Couplings predictions on amino acid preferences with H-CNN's predictions in Fig. 2.3F.

To measure the similarity between the predictions of H-CNN and EV Couplings, we define the profile overlap ρ for amino acid preferences at a given position. Specifically, for two sets of amino acid probabilities $P_A, P_B \in$ 19-simplex $\subset \mathbb{R}^{20}$, the vector overlap is defined as the normalized dot product between the centered probability vectors,

$$\rho = \frac{\big(P_A - \langle P_A\rangle_{\text{sites}}\big)\cdot\big(P_B - \langle P_B\rangle_{\text{sites}}\big)}{\sqrt{|P_A - \langle P_A\rangle_{\text{sites}}|^2 |P_B - \langle P_B\rangle_{\text{sites}}|^2}}. \tag{2.14}$$

Ultimately, we observed an average of 67% profile overlap. Furthermore, we observed that the distribution of overlapped vectors was clustered near 1 with a long tail towards 0 as shown in Fig. 2.3E. This distribution differed significantly from the distribution of overlaps obtained from a shuffling of the data over the sites.

Summary Overall, H-CNN is an efficient state-of-the-art predictor of amino acid identity based on local atomic surroundings with less training time than equally capable non-symmetry-aware predictors. These results highlight the utility of geometric deep learning's approach towards machine learning tasks on geometric data. H-CNN also stands out in the class of symmetry-aware models due to its superior classification ability than other symmetry-aware models despite using fewer parameters. Furthermore, H-CNN's predictions demonstrate that the network has learned features that are useful beyond the original amino acid classification task. H-CNN can use physicochemical information appropriately as evidenced by the physicochemical clustering of predictions as well as by the effect of ablation studies. H-CNN's predictions also quantitatively reflect the usage of different amino acids over evolutionary histories—a result that implies that H-CNN can predict the evolutionary fitness of mutations which is directly linked to a protein's quantitative function. These results, coupled with the promised interpretability and generalizability of equivariant neural networks, support our hypothesis that an equivariant neural network trained on amino acid classification could learn amino acid propensities that are useful for larger tasks of quantitative function prediction.

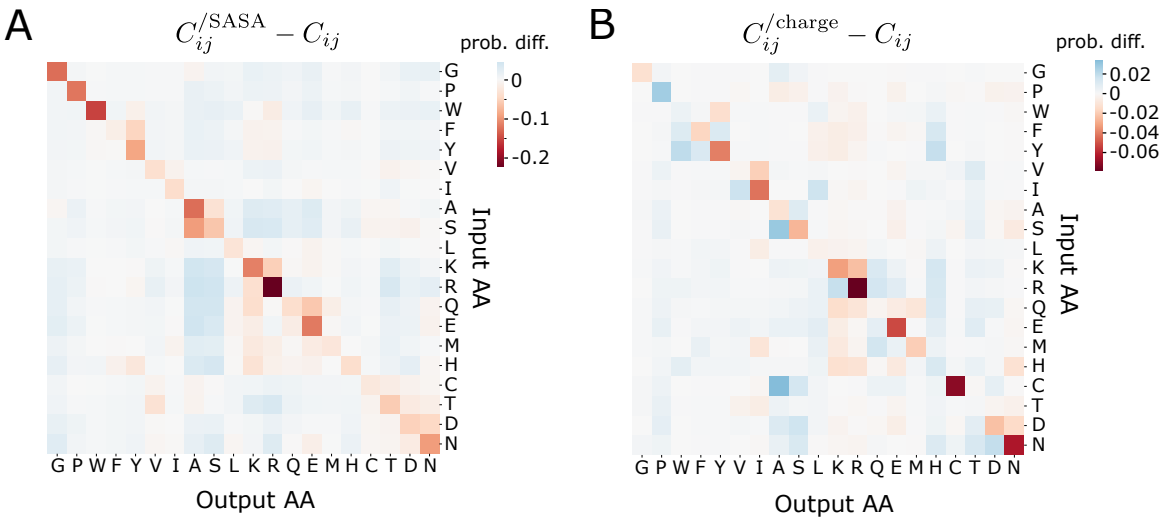


Figure 2.4: **H-CNN performance after charge and SASA ablation.** The difference in the confusion matrix for networks with ablated **(A)** SASA and **(B)** charge with respect to the network with the full information (Fig. 2.3A) is shown. Negative (red) values demonstrate less probability assigned to the input/predicted pair when the input is ablated, while positive (blue) values show increased probability mass under ablation. The removal of SASA (A) appears to decrease the network's ability to predict all amino acids as seen on the diagonal. The effect is most pronounced for hydrophobic amino acids in the upper left, with some hydrophilic amino acids (R, K, E) also displaying strong effects. Interestingly, most of the probability mass is on average redistributed to small amino acids A and S. When charge is removed (B), the network demonstrates worse predictions on charged and polar amino acids most notably R, C, N, and E. The overall network's accuracy after removing SASA and Charge are 57% and 56%, respectively.

Chapter 3

GENERALIZATION OF H-CNN AS A PHYSICAL POTENTIAL AND PREDICTOR OF PROTEIN FUNCTION

This chapter contains work from the following paper which is in submission at the time of writing:

> Pun, M. N., Ivanov, A., Bellamy, Q., Montague, Z., LaMont, C., Bradley, P., Otwinowski, J., & Nourmohammad, A. (2022). *Learning the shape of protein micro-environments with a holographic convolutional neural network* (p. 2022.10.31.514614). bioRxiv. https://doi.org/10.1101/2022.10.31.514614

In the last chapter, I have demonstrated that H-CNN can solve the amino acid classification task with state-of-the-art accuracy and the structure of its predictions reflects evolutionary usage of amino acids. This evolutionary usage of amino acids reflects the contributions of a particular amino acid to the fitness of a protein and subsequently to an organism. While how a protein contributes to the fitness of an organism is generally complex, it is intricately related to how well a protein performs its function. A polymerase that replicates DNA slowly may not allow its cells to reproduce and will confer low fitness to its organism. Meanwhile a polymerase that replicates DNA faster yet reliably will confer large fitness. Thus, the evolutionary use of amino acids is intricately related to the quantitative function of proteins and is in fact evidence that H-CNN can succeed at quantitative function prediction.

In this chapter, I will make the connection between H-CNN outputs and energetic contributions more rigorous focusing first on conceptual motivation and then on applications to relevant biological systems. In section 3.1, I will formalize how we can interpret H-CNN's output as approximations to free energies and I will demonstrate the physical reasonability of the energy interpretation of H-CNN by investigating how the H-CNN energetics respond to physical deformations of crystal structures. Then in sections 3.2 and 3.3, I will demonstrate H-

CNN's energetics correlate with experimentally measured intra-protein and protein-protein interaction energies in the systems of T4 Lysozyme and SARS-CoV-2 receptor binding domain-ACE2 complex.

3.1 H-CNN learns an effective physical potential for amino acids

The final step in determining the likelihoods of each amino acid according to H-CNN involves a softmax operation which maps a twenty-dimensional vector $\mathbf{E} \in \mathbb{R}^{20}$ to a probability distribution $\mathbf{P} \in$ 19-simplex $\subset \mathbb{R}^{20}$ via

$$P_\theta(\alpha) = \frac{e^{-\beta E^\alpha}}{\sum_\gamma e^{-\beta E^\gamma}} \tag{3.1}$$

where β is an inverse temperature that is chosen to be 1. It is well known that this operation is simply a Boltzmann distribution where the configurations in state space are the uses of each of the one amino acids and E^α are the energies of each configuration. Similarly, at equilibrium, single-mutant variants of a protein would be expected to be distributed according to the stabilities ΔG of the variants. Or, fixing an energetic scale by taking the wild-type sequence ΔG_{wt} as reference, the distribution of variants should follow

$$P(\alpha) = \frac{e^{-\beta \Delta\Delta G_\alpha}}{\sum_\gamma e^{-\beta \Delta\Delta G_\gamma}} \tag{3.2}$$

where $\Delta\Delta G_\alpha = \Delta G_\alpha - \Delta G_{\text{wt}}$. If our dataset of atomic environments generally spans the space of configurations and our model is regularized enough to prevent overfitting, we may assume that the distributions of amino acids in the context of certain atomic features are distributed in an equilibrium fashion. Although these conditions are not rigorously detailed here, following these assumptions, we may expect H-CNN pseudo-energies E_α to approximate physical free energies reasonably well and analyze H-CNN pseudo-energies under this hypothesis.

A map from atomic configurations to effective energy that respects the underlying symmetries of interactions should behave nicely in response to physical perturbations. Specifically, we hypothesize that if H-CNN is an effective physical potential for amino acids, then this potential should exhibit minima at crystal structure conformations that are locally quadratic with respect to changes to the structure. To test this hypothesis, we

characterize the response of the H-CNN predictions to physical distortions in native atomic micro-environments.

We introduce distortions through local *in silico* shear perturbation of the protein backbone at a given site i by an angle δ, resulting in a transformation of the backbone angles by $\phi_i \to \phi_i + \delta$, $\psi_{i-1} \to \psi_{i-1} - \delta$ (Fig. 3.1A). The shear perturbation is local with minimal downstream effects [22]. We perform this shearing for specific values of δ ranging from $-20°$ to $20°$. At each value of shear angle δ, we label the ultimate structure $\mathbf{x}_a(\delta)$ where a indexes each atom in the structure. Since the ultimate distortion to the protein structure will change based on the entire backbone geometry, we use the change in the pairwise distances to quantify the total distortion to the protein. Specifically, we measure the distortion of the protein structure due to shear by calculating the change in the root-mean-square (RMS) deviation in the pairwise distances of all atoms of the perturbed protein structure relative to that of the wild type (RMSΔD_{ab}, for all pairs of atoms (a, b)); Fig. 3.1B. The deviation of pairwise atomic distances is explicitly given by

$$\Delta D_{ab}(\delta) = |\mathbf{x}_a(\delta) - \mathbf{x}_b(\delta)| - |\mathbf{x}_a(0) - \mathbf{x}_b(0)|. \tag{3.3}$$

We then measure the response of H-CNN to this shear perturbation by analyzing the change in the pseudo-energies E_i^{α}. While it is tempting to expect that each pseudo-energy should follow our hypothesis above, we note that the perturbation also affects neighboring residues in 3D space and a perturbation that is favorable at one site may be unfavorable at all surrounding sites. Thus the magnitude of the energetic effect on the structures should be a function of the energy of all amino acids in the protein. To account for this distributed effect, we re-evaluate the pseudo-energy of each amino acid in the protein for a given distorted structure, and define the total H-CNN predicted energy by summing over the pseudo-energies of all the amino acids in a protein (Fig. 3.1C). The H-CNN defined energy of interest is

$$E(\delta) = \sum_{i \in \{1,\ldots,L\}} E_i^{\sigma_i}(\delta). \tag{3.4}$$

The change in the predicted energy of a protein due to distortion (relative to the wild-type structure) $\Delta E(\delta) = E(\delta)$–$E(0)$ is a measure of H-CNN's response to a given perturbation. A positive ΔE indicates an unfavorable change in the protein structure.

In protein G (PDB structure 1PGA), the change in the predicted energy ΔE as a function of distortion in the structure RMSΔD_{ab} due to shearing at different sites reveals two trends (Fig. 3.1D). First, the protein network energy appears to respond locally quadratically to perturbations. Second, perturbations generally result in higher protein network energy, corresponding to a less favorable protein micro-environment. Taken together, by training on a classification task and by constraining the network to respect the relevant rotational symmetry, H-CNN appears to have learned an effective physical potential for protein micro-environments in which the native crystal structure is generally more favorable and robust to local perturbations (i.e., it is at the energy minimum).

An alternative explanation for these minima is that data from the same distribution as the training data is generally more favorable regardless of the central amino acid. If this were true, then the energy of a random sequence $\hat{\sigma}$ should also lie at an energetic minima at $\delta = 0$ (albeit a minima that is absolutely less favorable than the true sequence σ). To investigate this, we generated random sequences $\hat{\sigma}$ and calculated energies using these random sequences. We sampled random sequences under two sampling schemes: (i) according to BLOSUM62 substitutability scores and (ii) according to uniform random distributions.

In Figure 3.2, we show a set of response curves for each of these sampling schemes. The appearance of a energy minima persists beyond the wild-type sequence in the case where biophysically similar amino acids are substituted in the energy sum (Fig. 3.2A). However, this pattern disappears when random amino acids are used to calculate the network energy (Fig. 3.2B). From this trend, it is clear that the network does not simply favor crystal structures in general. It only favors crystal structures when the true sequence or a physicochemically similar sequence is conditioned. This physically intuitive response of H-CNN to physical perturbations strengthens the case of using our network as an effective physical potential.

3.2 Prediction of stability of single point mutation variants of T4 lysozyme

Characterizing amino acid preferences in a protein neighborhood is closely related to the problem of finding the impact of mutations on protein function and indeed if H-CNN can be used to approximate physical energies in proteins, it should be capable of predicting

Figure 3.1: **Response of H-CNN predictions to physical distortions in a protein structure.** **(A)** The schematic shows shear perturbation in a protein backbone by an angle δ at site i as a rotation of side chains around the backbone by the angles $[\phi_i, \psi_{i-1}] \rightarrow [\phi_i + \delta, \psi_{i-1} - \delta]$ [22]. **(B)** Shearing changes the pairwise distance matrix D_{ab} between all atoms in a protein structure. The total physical distortion is computed as the root-mean-square of changes in the pairwise distances that are less than 10 Å (i.e., residues within the same neighborhood), multiplied by the sign of the change in the angle ψ. **(C)** For a given perturbation, the network energy E is determined by the sum of pseudo-energies of the wild-type amino acid at all sites in the protein, and the change in this quantity by shearing ΔE measures the tolerance of a structure to a given perturbation. **(D)** Panels show the change in the network energy in response to the structural distortion by shear perturbation at all sites in protein G, with the amino acid type and the site number indicated above each panel.

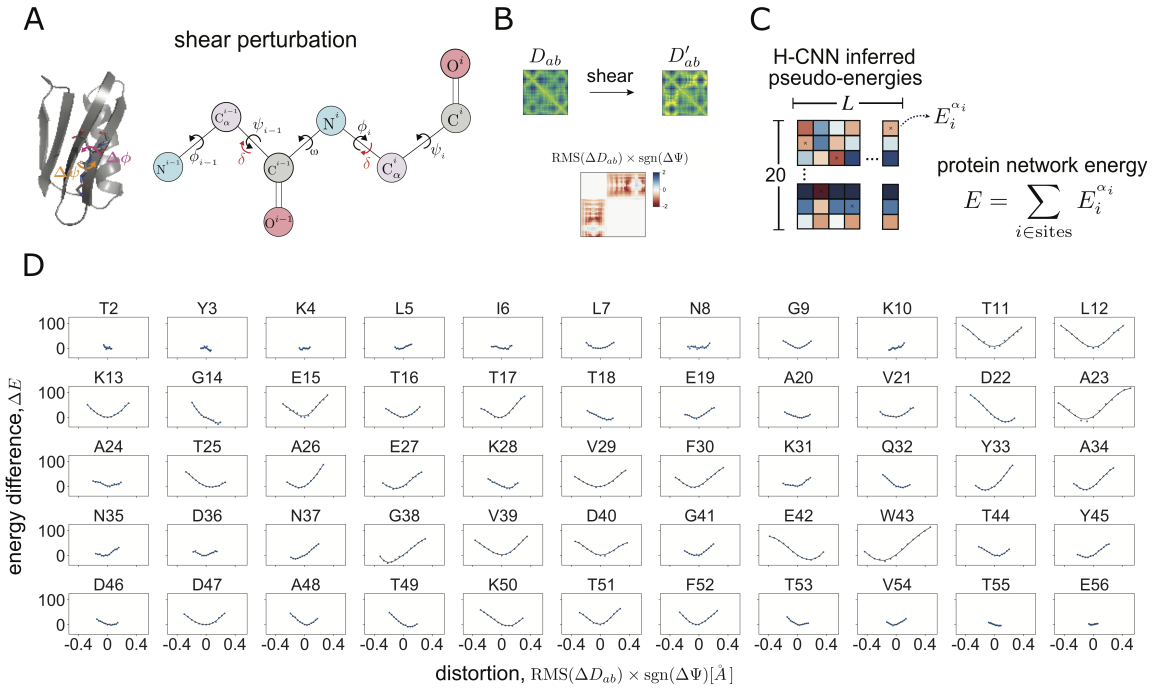
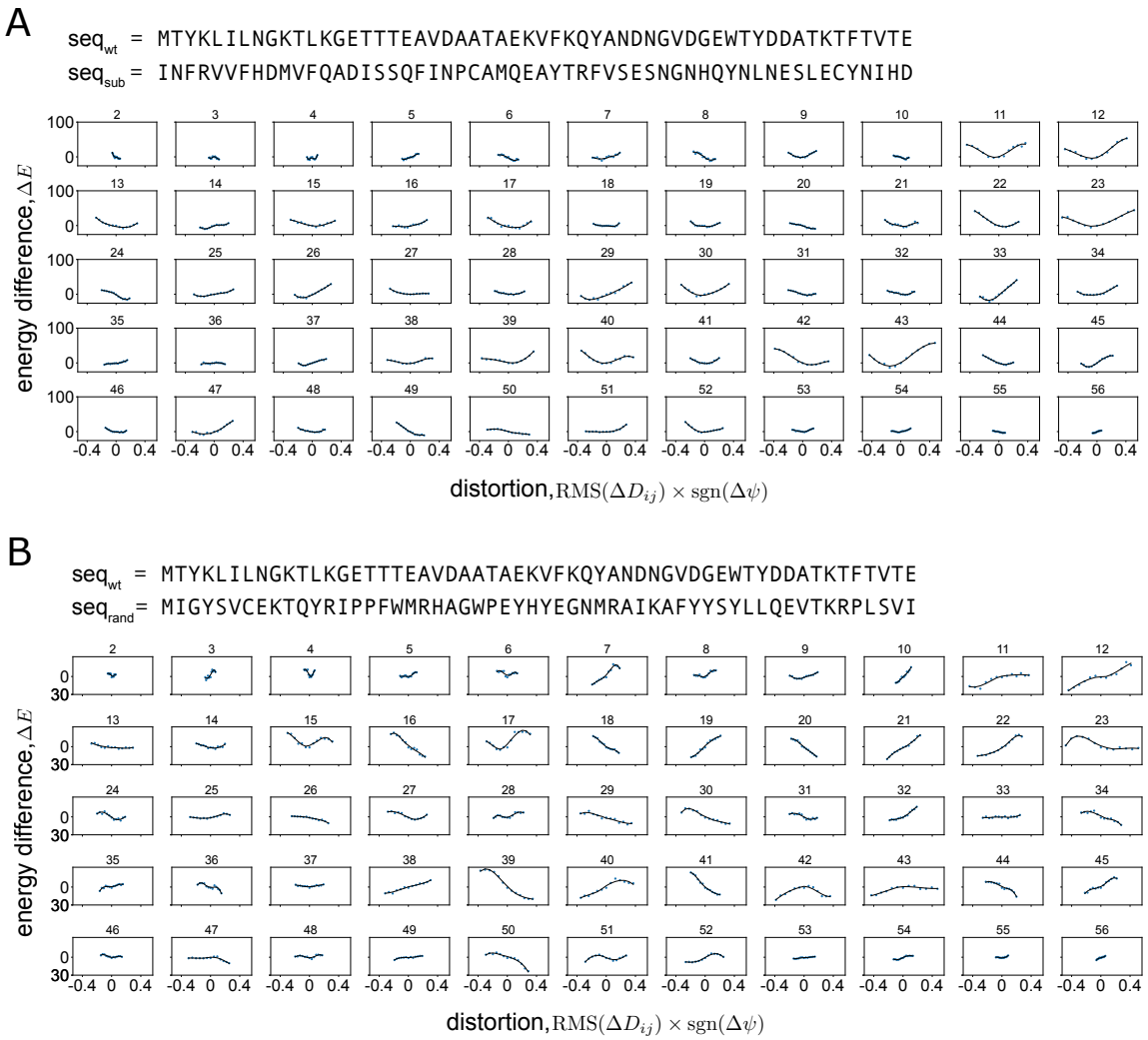

A
shear perturbation
B
shear
C
H-CNN inferred
pseudo-energies
20
protein network energy
D
energy difference, ΔE
distortion, $\mathrm{RMS}(\Delta D_{ab}) \times \mathrm{sgn}(\Delta\Psi)[\AA]$
T2 Y3 K4 L5 I6 L7 N8 G9 K10 T11 L12
K13 G14 E15 T16 T17 T18 E19 A20 V21 D22 A23
A24 T25 A26 E27 K28 V29 F30 K31 Q32 Y33 A34
N35 D36 N37 G38 V39 D40 G41 E42 W43 T44 Y45
D46 D47 A48 T49 K50 T51 F52 T53 V54 T55 E56

Figure 3.2: **Robustness of response to shear perturbation.** Similar to Fig. 3.1, but the change in network energy is evaluated for alternative amino acids at the shear position, while keeping the surrounding protein structure intact (i.e., using the wild-type neighborhood for model evaluation). The alternative amino acids are drawn according to the BLOSUM62 matrix in **(A)**, and randomly in **(B)**; the wild-type and the alternative sequence are shown each panel. When amino acids are substituted according to their BLOSUM62 scores, potential wells are generally recovered (A), while no energy minima is recovered for random substitutions (B). Specific sequences used are given by seq_{sub} (substitutions according to the BLOSUM62 matrix), and seq_{rand} (random substitutions) in each panel and are shown in comparison to the wild-type sequence seq_{wt}.


A
seq_wt = MTYKLILNGKTLKGETTTEAVDAATAEKVFKQYANDNGVDGEWTYDDATKTFTVTE
seq_sub = INFRVVFHDMVFQADISSQFINPCAMQEAYTRFVSESNGNHQYNLNESLECYNIHD
energy difference, ΔE
distortion, RMS(ΔD_ij) × sgn(Δψ)
B
seq_wt = MTYKLILNGKTLKGETTTEAVDAATAEKVFKQYANDNGVDGEWTYDDATKTFTVTE
seq_rand = MIGYSVCEKTQYRIPPFWMRHAGWPEYHYEGNMRAIKAFYYSYLLQEVTKRPLSVI
energy difference, ΔE
distortion, RMS(ΔD_ij) × sgn(Δψ)

experimentally measured stabilities. Here, we test the accuracy of H-CNN in predicting the stability effect of mutations in 40 different variants of the T4 lysozyme protein. Each of these variants is one amino acid away from the wild type, with variations spanning 23 residues of the protein. Notably, the tertiary structure of the wild-type T4 lysozyme protein as well as the 40 mutants are available through different studies [50, 49, 59, 89, 33, 145, 30, 84, 79, 6, 157, 159, 90, 95, 94, 43, 102, 137]; see Table B.1 for details on these mutants.

H-CNN predicts that the wild-type amino acids are the most favorable in the wild-type structure, while the mutant amino acids are generally more favorable in the mutant structures, regardless of their stabilizing effects (Fig. 3.3A). These variant-specific preferences are not surprising since the folded protein structure can relax to accommodate for amino acid changes, resulting in a structural neighborhood that is more consistent with the statistics of the micro-environments around the mutated amino acid than that of the wild type. However, the confidence that H-CNN has in associating an amino acid with a given structural neighborhood can change depending on the stability effect of the mutation. The log-ratio of the H-CNN inferred probability for the mutant amino acid in the mutant structure versus that of the wild-type amino acid in the wild-type structure $\Delta \log P = \log P_{\text{mut}}/P_{\text{wt}}$ can provide an approximation to the change in energy of formation $\Delta\Delta G$ associated with the stability of a mutation.

The inferred H-CNN predicted log-probability ratio is generally negative for destabilizing mutations, and nonnegative for neutral/weakly beneficial mutations (Fig. 3.3B). Previously, a structure-based CNN model with voxelized protein structures has shown a similar qualitative result [137]. Further quantitative analysis shows that the log-probability ratio is 67% correlated with the experimentally evaluated $\Delta\Delta G$ values for these variants (Fig 3.3C). Moreover, the receiver-operating-characteristic (ROC) curves in Fig. 3.3D show that the log-ratio of amino acid probabilities can reliably discriminate between destabilizing and neutral mutations, with an area under the curve (AUC) of 0.90.

The availability of tertiary structures for a large number of variants is a unique feature of this dataset, and in most cases such structural resolution is not accessible. To overcome this limitation and predict the stability effect of mutations by relying on the wild-type structure alone, we used PyRosetta to relax the wild-type T4 lysozyme structure around a specified

amino acid change [22]. We find that the log-probability ratios $\Delta \log P_{(\mathrm{sil})}$ estimated based on these *in silico* relaxed mutant structures are mostly negative (nonnegative) for destabilizing (neutral) mutations (Fig. 3.4C) and are correlated with the stability effect of mutations $\Delta\Delta G$ (Fig. 3.4G). However, structural relaxation can add noise to the data, causing the protein micro-environments to deviate from the natural structures that H-CNN is trained on. Thus, using the *in silico* relaxed structures slightly reduces the discrimination power of our model between deleterious and near-neutral mutations (AUC = 0.83); see Fig. 3.3D.

In contrast, the preferences estimated based on the wild-type structure only can discriminate between destabilizing and neutral mutations very well, even though most mutations are inferred to be deleterious with respect to the wild type (AUC = 0.93 in Figs. 3.3D & 3.4A). In other words, by using the wild-type structure only, our model can predict the relative stability effect of mutations correctly but not the sign of $\Delta\Delta G$ (Fig. 3.4). Indeed, our inferred log-probability ratios based on the wild-type structure show a substantial correlation of 64% (Pearson correlation) with the stability effect of a much larger set of 310 single point mutants [131], for which protein structures are not available (Fig. 3.5A).

When no experimentally determined structure is available, computationally resolved protein structures from AlphaFold can also be used to predict the stability effect of mutations. The H-CNN predictions using the template-free AlphaFold2 predicted structure of T4 lysozyme wild-type sequence display substantial discrimination ability between destabilizing and near-neutral mutations (Fig. 3.3D) and are correlated with the mutants' $\Delta\Delta G$ values (Figs. 3.4D,H & 3.5B).

In summary, H-CNN predictions can be used to predict relative stability effects of single point mutations as measured in laboratory conditions. We now will provide further evidence of H-CNN as an effective potential for amino acids by demonstrating its ability to predict relative effects of mutations in the case of protein-protein interactions.

3.3 Prediction of SARS-CoV-2 RBD binding to ACE2 receptor

Recent deep mutational scanning (DMS) experiments measured the effect of thousands of mutations in the receptor-binding domain (RBD) of SARS-CoV-2 on the folding of the RBD (through expression measurements) and its binding to the human Angiotensin-Converting

Figure 3.3: **Predicting the stability effect of mutations in T4 lysozyme with H-CNN. (A)** Heatmaps of H-CNN predicted log probability of different amino acids (columns) relative to that of the wild-type amino acid for 40 variants with single amino acid substitution from the wild type (rows). For each variant (row), the position and the identity of the wild-type amino acid and the mutation are denoted between the two heatmaps as: wild type, site number, mutation. The left panel shows the predictions using the wild-type protein structure (subscript (wt)), while the right panel shows the predictions using the structure of the specified mutant at each row (subscript (mut)). In each row the wild-type amino acid is indicated by an $\times$, and a dotted box shows the amino acid of the mutant. **(B)** The H-CNN predicted log-probability ratio $\Delta \log P = \log P^{\text{mut}}_{(\text{mut})}/P^{\text{wt}}_{(\text{wt})}$ of the mutant amino acid on the mutant structure with respect to the wild-type amino acid on the wild-type structure is shown for all variants. The predicted ratios for destabilizing mutations are negative, while those for the neutral / beneficial mutations are positive. **(C)** The H-CNN predicted log-probability ratio $\Delta \log P$ shown against the experimentally evaluated $\Delta\Delta G$ for the stability effect of mutations in each protein structure; Pearson correlation of 67%. **(D)** The true positive vs. false positive rates (ROC curve) is shown for classification of the 40 amino acid mutations in (A) into deleterious and neutral/beneficial classes based on the H-CNN predicted log-probability ratios $\Delta \log P$, using the wild-type and the mutant structures (orange), only the wild-type structure (purple), the wild-type structure for the wild type and an *in silico* relaxed structure for each mutant (green), and the AlphaFold predicted structure of the wild type (blue). The corresponding area under the curves (AUCs) are reported in the figure.

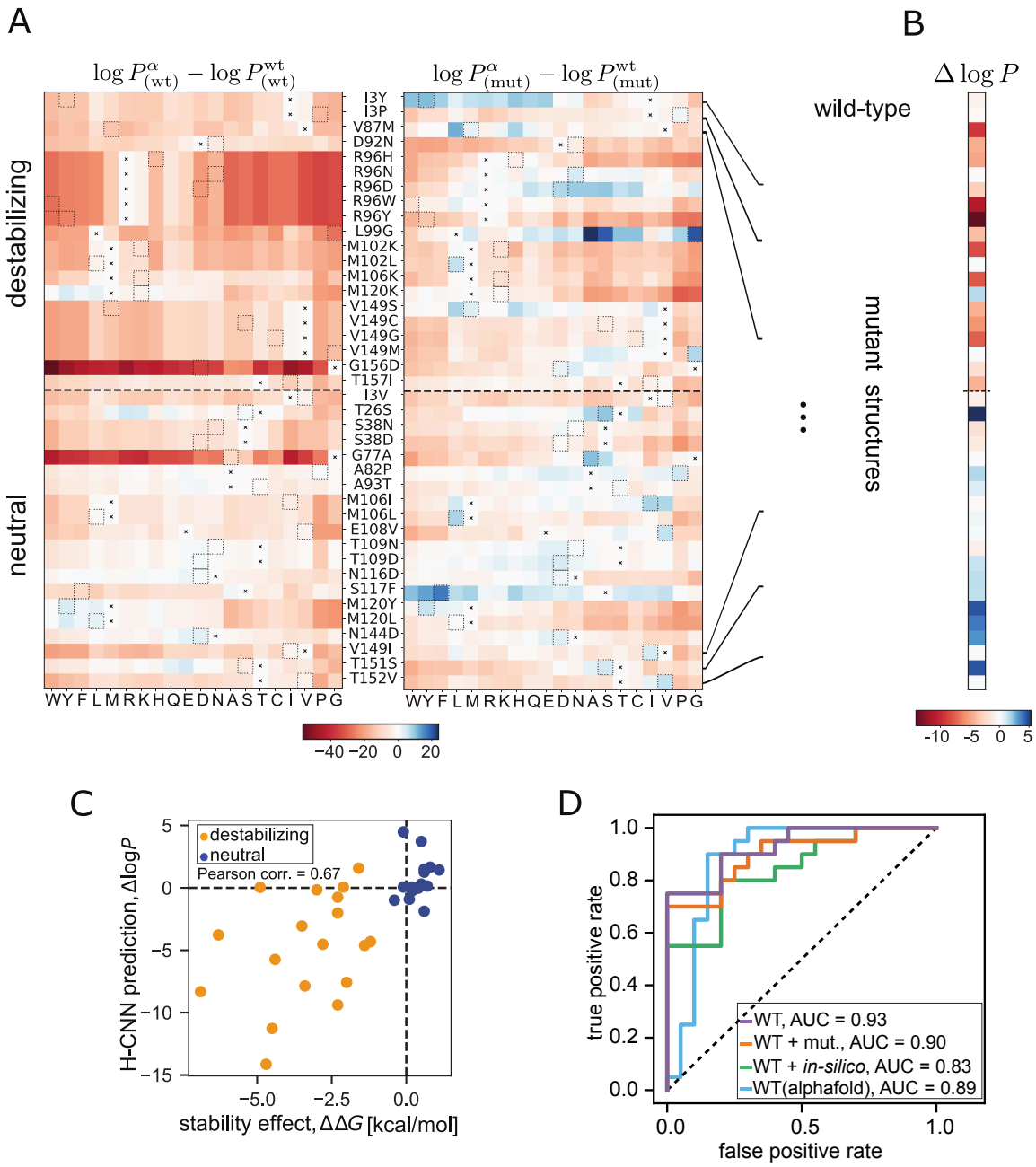
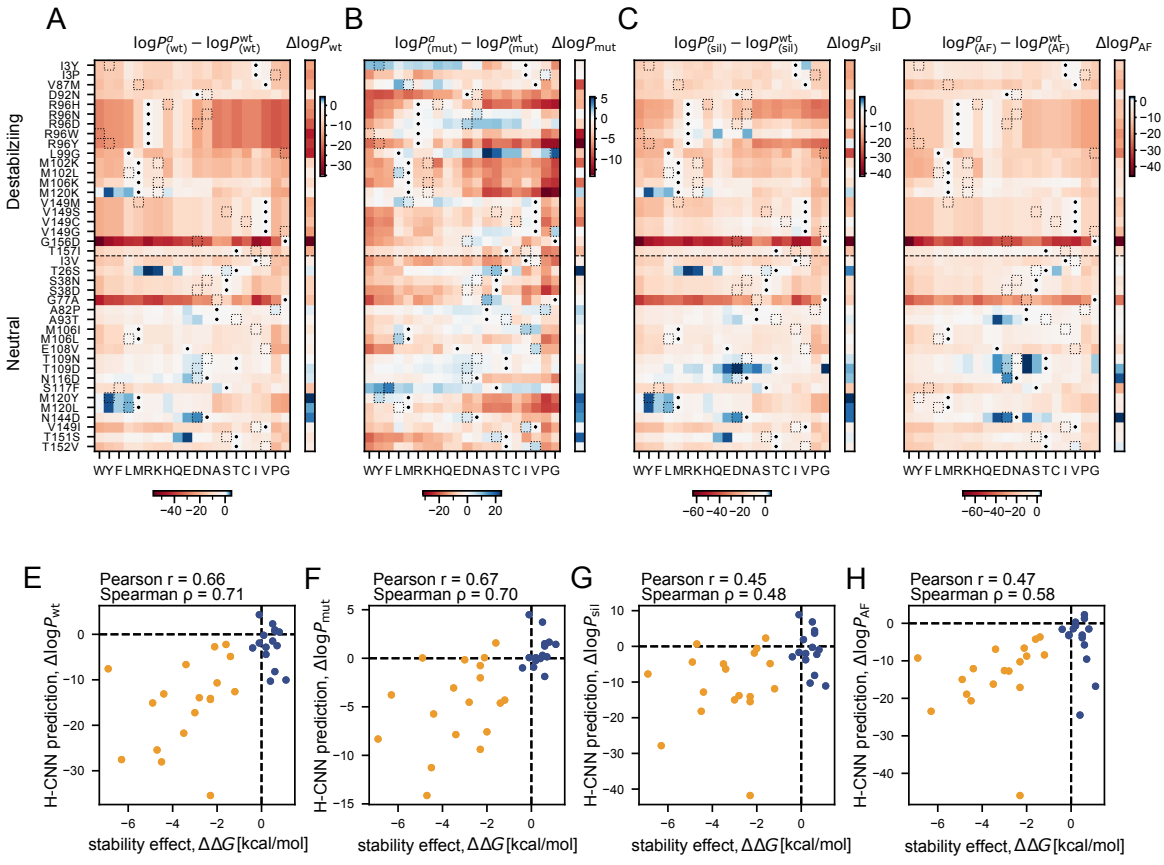

A
B
C
D
destabilizing
neutral
wild-type
mutant structures
Δ log P
I3Y
I3P
V87M
D92N
R96H
R96N
R96D
R96W
R96Y
L99G
M102K
M102L
M106K
M120K
V149S
V149C
V149G
V149M
G156D
T157I
I3V
T26S
S38N
S38D
G77A
A82P
A93T
M106I
M106L
E108V
T109N
T109D
N116D
S117F
M120Y
M120L
N144D
V149I
T151S
T152V
WYFLMRKHQEDNASTCIVPG
destabilizing
neutral
Pearson corr. = 0.67
H-CNN prediction, ΔlogP
stability effect, ΔΔG [kcal/mol]
true positive rate
false positive rate
WT, AUC = 0.93
WT + mut., AUC = 0.90
WT + in-silico, AUC = 0.83
WT(alphafold), AUC = 0.89

Figure 3.4: **Predictions for the stability effect of mutations in T4 lysozyme with different protein structures.** Similar to Fig. 3.3A, in each heatmap, the dots indicate the identity of the wild-type amino acid, and the dotted squares indicate the amino acid in the specified position for the mutated variant in each row. The predictions for log-probabilities are evaluated using the wild-type structure (A), the mutant crystal structures (B), the wild-type structure relaxed *in silico* via PyRosetta for the specified substitution (C), and the computationally predicted structure of the wild-type T4 lysozyme using AlphaFold (D). The relative log-probabilities of the mutations of interest are shown separately in one column next to each heatmap. They are calculated with the structural information that is available in different scenarios. For the wild-type, mutant, and *in silico* structures, $\Delta \log P_* = \log P^{\text{mut}}_{(*)}/P^{\text{wt}}_{(\text{wt})}$, with $*$ indicating the specified protein structure in each case. This definition is akin to $\Delta\Delta G$. For the AlphaFold structure, we use $\Delta \log P_{\text{AF}} = \log P^{\text{mut}}_{(\text{AF})}/P^{\text{wt}}_{(\text{AF})}$—a quantity which could be evaluated in the absence of any experimental structural knowledge for a protein. **(E-H)** The predicted relative log-probabilities shown in panels (A-D) are plotted against the experimentally measured $\Delta\Delta G$ values for each variant. Destabilizing mutations are shown in orange and neutral/beneficial mutations are shown in blue. Overall, H-CNN predictions correlate well with the mutational effects on protein stability, with the Pearson and Spearman correlations indicated in each panel and AUC values to discriminate between destabilizing and neutral mutations indicated in Fig. 3.3D. Accounting for changes in the local protein structures due to mutations in (B, C) sets the correct reference point in predictions such that the estimated log-probability ratio is overall positive for neutral/beneficial mutations and negative for destabilizing mutations.


A
B
C
D
Destabilizing
Neutral
E
F
G
H
Pearson r = 0.66
Spearman ρ = 0.71
Pearson r = 0.67
Spearman ρ = 0.70
Pearson r = 0.45
Spearman ρ = 0.48
Pearson r = 0.47
Spearman ρ = 0.58
stability effect, ΔΔG [kcal/mol]

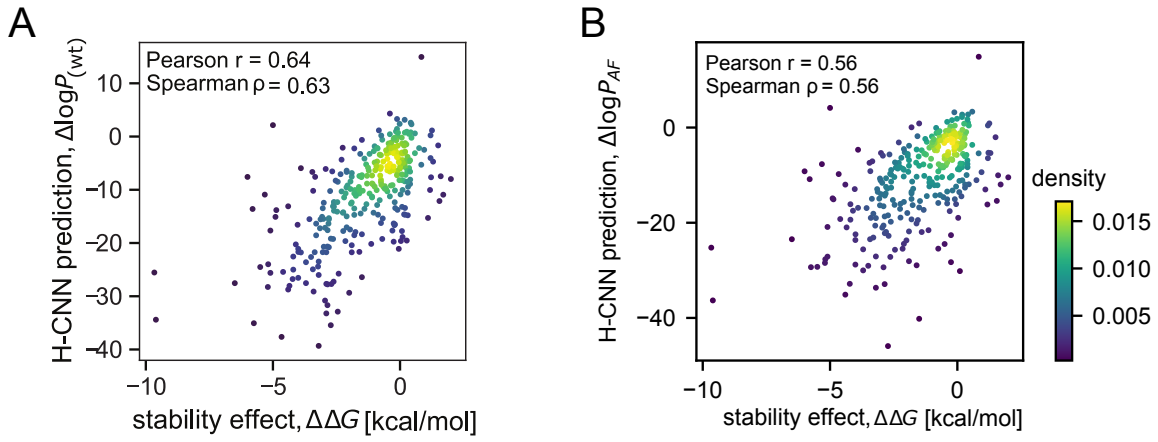


Figure 3.5: **H-CNN predictions for the stability effect of all available single point mutations in T4 lysozyme.** H-CNN predictions for the relative log-probabilities using **(A)** the wild-type structure $\Delta \log P_{\mathrm{wt}}$ and **(B)** the AlphaFold predicted structure $\Delta \log P_{\mathrm{AF}}$ are shown against the experimentally measured $\Delta\Delta G$ values for 310 single point mutation variants of T4 lysozyme. Mean $\Delta\Delta G$ was used when multiple experiments reported values for the same variant. The colors show the density of points in each panel as calculated via Gaussian kernel density estimation. H-CNN predictions correlate well with the experimental values, with the Pearson and Spearman correlations indicated in each panel.

Enzyme 2 (ACE2) receptor [125, 124].

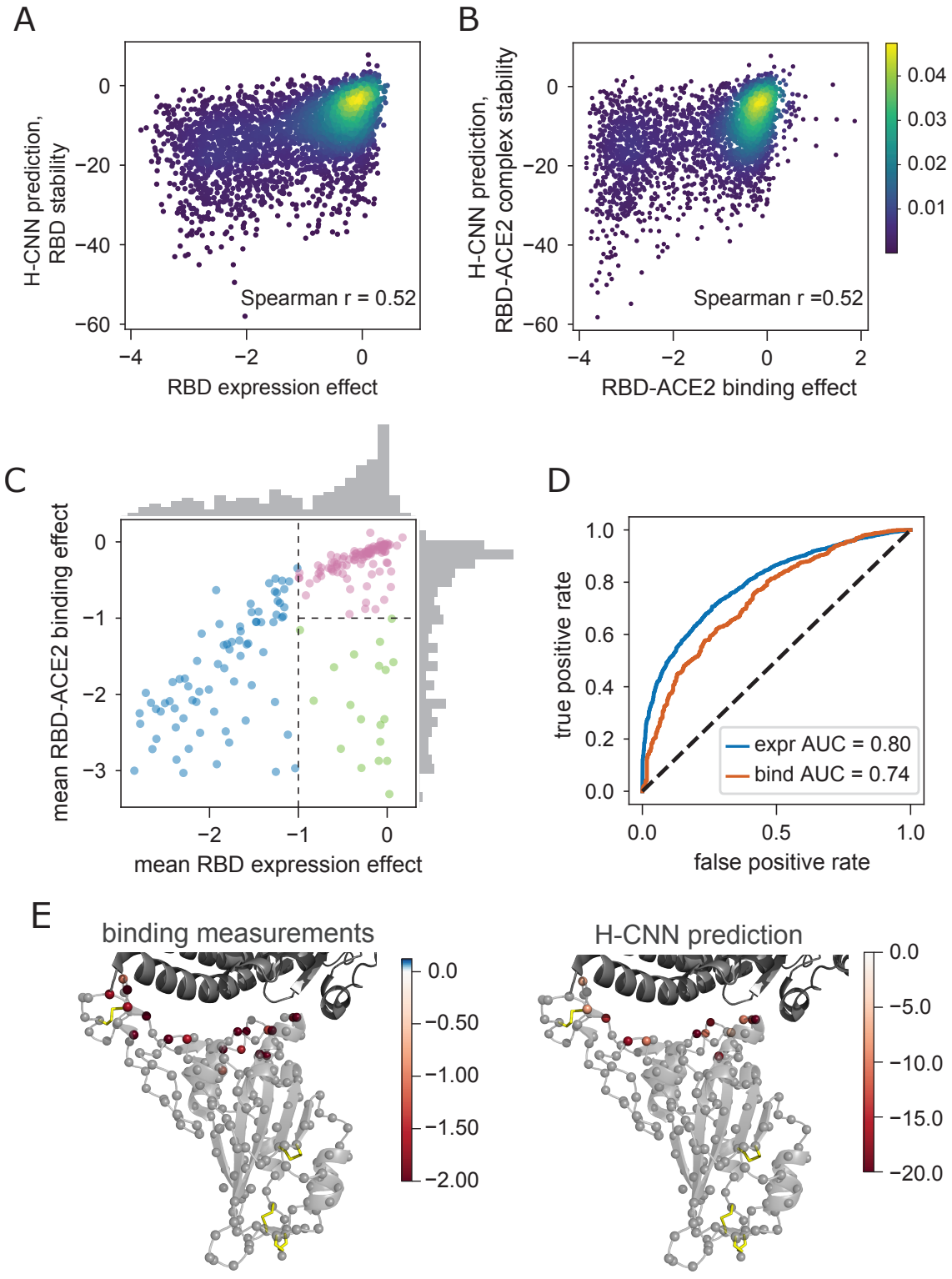
H-CNN can be used to predict the effect of mutations on RBD, either in isolation or bound to the ACE2 receptor. The former can be interpreted as the effect of mutations on the stability of RBD, which is measured by the expression of the folded domain in the experiments [2, 125, 124], while the latter can be used to characterize amino acid preferences for binding at the RBD-ACE2 interface. Fig. 3.6A,B shows that the H-CNN predictions are correlated with the stability and binding measurements in the DMS experiments from ref. [124]; site specific effects are depicted in Figs. B.1 & B.2.

The average effect of mutations on expression and binding can define three categories of sites and/or mutations (Fig. 3.6C): (i) sites that are intolerant to mutations (due to destabilizing effects) and show a substantially reduced expression of mutants (blue), (ii) sites that are tolerant of mutations for expression but not binding (green), and (iii) sites that are tolerant of mutations for both expression and binding (pink). Using the isolated structure of RBD, H-CNN can well classify mutations according to their stability effect (AUC = 0.8; Fig. 3.6D). Similarly, with the structure of the RBD-ACE2 complex, H-CNN can classify mutations according to their tolerance for binding (AUC = 0.74; Fig. 3.6D).

Expectedly, the sites that are tolerant of mutations for expression but not binding (green category from the DMS data in Fig. 3.6C) are located at the interface of the RBD-ACE2 complex, and H-CNN correctly predicts this composition (Fig. 3.6E & B.2). The overall impact of mutations on binding for these sites is shown in Fig. 3.6E.

Identifying candidate sites that can tolerate mutations and can potentially improve binding is important for designing targeted mutagenesis experiments. Instead of agnostically scanning single point and (a few) double mutations over all sites, these predictions can inform experiments to preferentially scan combinations of viable mutations on a smaller set of candidate sites. In previous work, evolutionary information was used to design such targeted mutagenesis for the HA and NA proteins of influenza [143, 158]. A principled structure-based model could substantially improve the design of these experiments.

Figure 3.6: **Predicting the stability and binding of the RBD protein of SARS-CoV-2 with H-CNN. (A)** The density plot shows H-CNN predictions for the RBD stability, using the isolated protein structure of RBD, against the mutational effects on the RBD expression from the DMS experiments; Spearman correlation $r = 0.52$. **(B)** The density plot shows H-CNN predictions for the RBD binding to the ACE2 receptor, using the co-crystallized RBD-ACE2 protein structure, against the DMS measurements for mutational effects on binding; shared color bar for (A) and (B). **(C)** The mean effect of mutations at each site on the RBD-ACE2 binding is shown against the mean effect on the RBD expression. The histograms show the corresponding distribution of effects across sites along each axis. The categories are shown: (i) sites that are intolerant to mutations due to destabilizing effect, i.e., low expression (blue), (ii) sites that are tolerant of mutations for expression but not binding (green), and (iii) sites that are tolerant of mutations for both expression and binding (pink). **(D)** Blue: true positive vs. false positive rate (ROC curve) for classification of amino acid mutations into stable (expr > -1) vs. unstable (expr < -1), based on the H-CNN predictions using the isolated RBD structure; AUC = 0.8. Red: the ROC curve for mutation classification into bound (bind > -1) vs. unbound (bind < -1), based on the H-CNN predictions using the co-crystallized RBD-ACE2 structure; AUC = 0.74. **(E)** The effect of mutations on binding from the DMS experimental data for the green sites in (C) (top) and the corresponding H-CNN predictions from the RBD-ACE2 structure complex for sites identified by H-CNN in Fig. 3.7 to be tolerant of mutations for stability but not binding (bottom) are shown throughout the structure.


A
H-CNN prediction, RBD stability
RBD expression effect
Spearman r = 0.52
B
H-CNN prediction, RBD-ACE2 complex stability
RBD-ACE2 binding effect
Spearman r =0.52
C
mean RBD-ACE2 binding effect
mean RBD expression effect
D
true positive rate
false positive rate
expr AUC = 0.80
bind AUC = 0.74
E
binding measurements
H-CNN prediction

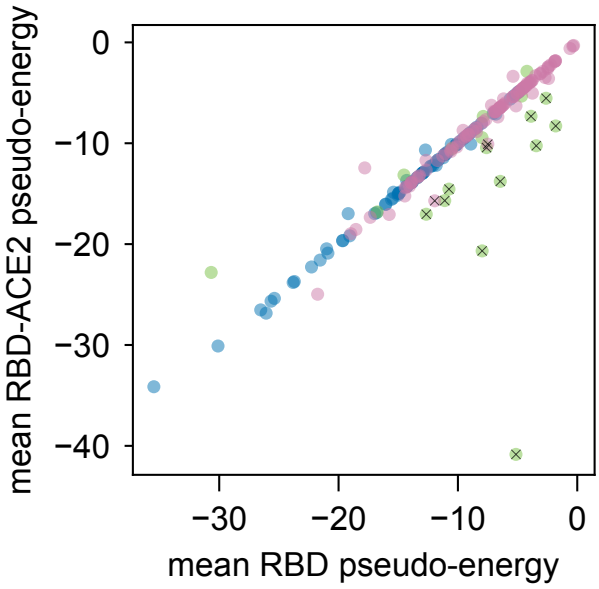


Figure 3.7: **H-CNN predictions for stability and binding of RBD.** The H-CNN predicted mean pseudo-energy per site predicted from the RBD-ACE2 protein complex (measure of binding) is plotted as a function of the mean pseudo-energy per site, inferred from the isolated RBD structure (measure of stability). Points are colored according to their experimental values of expression and binding as illustrated in Fig. 3.6. Most points lie on the diagonal since the RBD pseudo-energy is determined from the RBD-ACE2 crystal structure with the ACE2 masked. However, sites at the RBD-ACE2 interface show deviations from the diagonal. H-CNN prediction for sites that are tolerant of mutations for stability but not binding are indicated by crosses (below the diagonal points with large stability values). These sites tend to be at the RBD-ACE2 interface and overlap with the green category from Fig. 3.6.

Chapter 4

INTERROGATING T CELL RECEPTOR SPECIFICITY WITH H-CNN

In the previous chapter, I proposed an interpretation of H-CNN pseudo-energies as approximations to free energies of amino acids in atomic environments. I demonstrated that the pseudo-energies are consistent with this interpretation due to their physical response to structural perturbations and their ability to rank lab-measured intra-protein and protein-protein interaction energies in T4 lysozyme and the SARS-CoV-2 RBD-ACE2 complex respectively.

Having demonstrated the zero-shot generalization capabilities of H-CNN, I will now focus on a notoriously challenging protein, the T cell receptor (TCR). Since this protein has not yet been introduced, I will begin in section 4.1 with a brief introduction to T cells in the context of the adaptive immune system, the role of TCRs in immunity, and broader motivation for studying TCR specificity. I will also introduce a property known as the "specificity" of a TCR and cover the current state-of-the-art methods for predicting this property. Finally, I will make a case for why H-CNN offers a promising orthogonal approach to this prediction task and then outline the rest of the scientific results presented in this chapter.

4.1 Motivation

4.1.1 T cells and the adaptive immune system

T cells are a type of white blood cell that play a critical role in the adaptive immune system's response to pathogens. Through a process known as antigen presentation, all nucleated cells will signal their identity by presenting fragments of their internal proteins called peptides on surface proteins called Major Histocompatibility Complexes (MHCs) [101]. T cells are tasked with scanning these peptide-MHC complexes (pMHCs) via their own surface proteins called T cell receptors (TCRs). T cells are trained to detect foreign peptides and take action

when they encounter them. For example, killer T cells will respond to foreign peptides by inducing death in the cell presenting them. Meanwhile helper T cells will recruit other arms of the immune system such as B cells to mount their own response. In humans and other vertebrates, T cells can stop the spread of viral infections by mounting a response against the infected cells.

While the ultimate immune response depends on the T cell type, TCR binding to antigen is a necessary condition for any T-cell-mediated response [144]. The peptide fragments to which a given TCR will bind is called an epitope, and the immune system is able to protect against diverse sets of pathogens by utilizing many TCRs each of which can bind to many different epitopes [156, 13, 38, 28, 153, 62]. In fact, one TCR can bind to as many as one million different peptides [155].

A key feature of TCR binding is the low-affinity nature of epitope binding. In contrast to antibodies which are selected for strong binding to a particular pathogen, properly functioning T cells cannot bear TCRs that bind strongly to self peptides. To prevent this unwanted binding, TCRs that bind strongly to self are eliminated from the population in a process known as negative selection [116, 103]. Ultimately a healthy T cell population will feature TCRs that do not bind strongly to self-peptides but bind strongly enough to pathogenic peptides to effectively signal viral presence to the immune system at-large.

The low-affinity and degenerate nature of TCR binding makes prediction of TCR specificity, or the set of peptides to which a TCR would bind to, a difficult problem. Despite this challenge, understanding and prediction of TCR specificity would have significant impact in the fields of immunology and medicine. Lab-engineered receptors known as chimeric antigen receptors (CARs) displayed on CAR T cells [113] are the current state-of-the-art cancer immunotherapy [127, 4]. Prediction of TCR specificity would help guide design of future therapeutic receptors. Furthermore, understanding disease specific TCR specificity could shed light on the nature of auto-immune disease. And finally, characterization of TCR specificity could offer insights into fundamental questions in immunology such how cross reactive are TCRs in general.

The current state-of-the-art of TCR specificity prediction methods are indeed rooted in ML. However, due to the relative abundance of sequence data compared to structural

data, most methods are predominantly, if not exclusively, sequence-based [40, 45, 46, 64, 92, 118, 147]. While these models do achieve some success at predicting TCR specificity, recent results have demonstrated that structural features offer better predictions of quantitative affinities [88]. At the time of writing there are just over 327 co-complexed TCR-pMHC structures [47] with a heavy bias towards well-studied systems. Considering the scarcity of TCR-pMHC structural data in the context of diversity of the three interacting proteins, structure-based methods cannot train on TCR-pMHC data directly without severe risk of overfitting. Currently, structure-based methods largely use biochemistry-informed features. To carry out a richer data-driven structural analysis, ML methods for TCR specificity prediction must rely on learning generalizable features from the larger body of protein structures.

Following the demonstrated generalization of H-CNN in zero-shot quantitative function prediction tasks, I propose H-CNN as a robust and efficient model for predicting TCR specificity from a structural perspective. In the rest of this chapter, I will present H-CNN's ability to predict TCR specificity in a number of settings. First in section 4.3, I demonstrate once again H-CNN's zero-shot predictive power over TCR-pMHC binding affinity. Then in section 4.4, I examine H-CNN's ability to score TCR-pMHC docked structures. Finally, in section 4.5, I utilize H-CNN in a generative framework to design peptides that bind to given TCRs.

Before embarking on this analysis, I shall review some preliminary details on TCR-pMHC structures and notation conventions used that will be necessary for the discussion of results later.

4.2 Preliminaries on TCR structure

4.2.1 TCR structure and function

TCRs consist of two proteins chains α and β which together form one heterodimeric complex. The complex is generally broken down into three regions: (i) the transmembrane region which is embedded in the cell membrane, (ii) the constant region which is extracellular but closest to the membrane, and (iii) the variable region which is furthest from the membrane

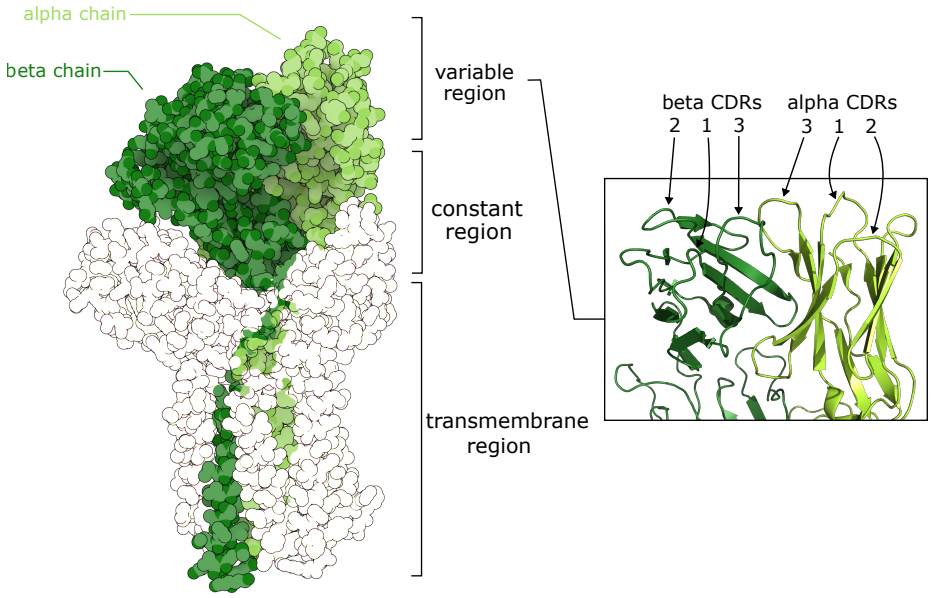


Figure 4.1: **Overview of the structure of a T cell receptor.** A TCR, as it is found on the surface of T cells, is shown in green with co-complexed CD3 proteins shown in white with black outlines. The alpha and beta chains are shown in light and dark green respectively. The variable, constant, and transmembrane regions are labeled and a detailed view of the variable region is shown at right. CDR loops are indicated with arrows.

and binds to the pMHC; see Fig. 4.1 for details. The variable region is the substructure that interacts with and binds to pMHC and is most important for determining TCR specificity.

The variable region is made up of six loops termed the complementarity determining regions (CDRs). All six regions are observed to contribute to TCR-pMHC interactions to varying degrees with the β CDR3 displaying the most contacts with the peptides on average [15, 72, 112]. Notably the β CDR3 also shows the largest amount of sequence variability due to diversity created in a genetic process called VDJ recombination.

4.2.2 Notation

We will use the following notation to describe TCR-pMHC complexes. Following the literature, we will refer to sites in the peptide of a TCR-pMHC complex as p1, p2, p3 etc. A

site with a specific amino acid will feature the one letter abbreviation of that amino acid after the site. For example, a glycine at position 4 will be noted p4G. All sites in the MHC and TCR will be denoted by the amino acid and the site number with a reference to the protein. For example, when referring to an asparagine at site 97 of a TCR β chain, we may refer to this as N97 in the TCR β chain or more specifically N97 in the βCDR3.

MHC molecules are diverse proteins with many alleles found both across and within organisms. Often times we will refer to a given allele by the allele name, such as A*02:01. Furthermore, in the context of humans, MHC molecules are referred to as human leukocyte antigen (HLA) and thus we will often refer to specific allele studied predominantly in this thesis HLA-A*02:01.

4.3 TCR-pMHC K_d prediction

4.3.1 Experimental measurement of K_d

TCR-pMHC binding is generally treated as a reversible process described by the chemical reaction

$$\text{TCR} + \text{pMHC} \underset{k_{\text{off}}}{\overset{k_{\text{on}}}{\rightleftharpoons}} \text{TCR:pMHC} \tag{4.1}$$

where k_{on} is the second order rate of complex formation and k_{off} is the first order rate of complex dissociation [129]. In general, these rates are used to calculate the dissociation constant of the reaction $K_d = k_{\text{off}}/k_{\text{on}}$ which has dimensions of concentration. Specifically, K_d describes the concentration of TCR at which half of the pMHC molecules will be associated with TCR. A smaller value of K_d indicates a higher affinity. For natural TCRs, K_d is usually in the range of 1-100 μM [31, 67, 166]. On the other hand, engineered TCRs can have nanomolar affinities [123] and have even been observed to exhibit affinities as strong as 26 pM [130, 76, 35].

Given a specific TCR-pMHC system, K_d is most accurately measured via surface plasmon resonance (SPR) spectroscopy experiments whereby the TCR (or MHC) is bound to a conducting plate and solvated MHC (or TCR) is introduced and allowed to bind to the protein attached to the plate.[1] This binding modulates the index of refraction near the

[1]For an overview of SPR spectroscopy as a tool for studying protein-protein interactions in general see [96]

MHC	TCR	peptide	N	measurement	source
A*02	A6	Tax	16	K_d	[99]
	1G4	NY-ESO-1	9		
			7	3d/2d affinity	[163]
			11	K_d	[3]
	DMF5	MART-1	9		[104]

Table 4.1: **Summary of TCR-pMHC K_d measurements with accompanying structures.**

plate which directly influences the amount of resonance observed. We have collected a small dataset of SPR-measured affinities [163, 104, 99, 3] for which a crystal structure of the complex with at least one peptide exists [27, 21, 42, 23] to the best of the author's knowledge.

4.3.2 TCR-pMHC systems analyzed

The collected dataset of paired K_d measurements and crystal-structure templates involves three TCRs binding to peptides presented by the human MHC allele A*02:01.[2] A summary of these data can be seen in table 4.1. Before presenting the predictions, I will briefly provide context on the systems.

DMF5 binding to MART-like peptides

One of the first TCRs used in cancer gene therapy [18], DMF5 binds to the melanoma tumor antigen MART-1.[3] Some cancer patients develop natural immune responses to tumor cells, and MART-1 is an example of a peptide involved in natural immune recognition of human

and for a review focused specifically on measuring TCR-pMHC interactions via SPR see [100].

[2] A*02:01 is the most represented MHC allele in the structural database TCR3d as it is present in 92 of the 327 complexed structures [47].

[3] Short for Melanoma Antigen Recognized by T cells and also known as Melan-A.

melanoma [108, 111]. DMF5 is an engineered receptor designed specifically to recognize MART-1 [63]. Notably it binds MART-1-like peptides that take two conformations [104]. Binding affinity of DMF5 to MART-1-like peptides was measured in [104].

1G4 binding to NY-ESO-1

New York esophageal squamous cell carcinoma 1 (NY-ESO-1) is a cancer-testis antigen that is expressed in many different types of cancer and has been studied as a target for cancer immunotherapies [37]. The TCR 1G4 binds to NY-ESO-1 peptides, and binding affinities of 1G4 to NY-ESO-1-like peptides were measured in [163, 3, 99].

A6 binding to Tax

The Tax protein of the human T-lymphotrophic virus type 1 (HTLV-1) is recognized by T cells in HTLV-1-infected patients [107, 139]. Binding affinities of Tax peptides were measured in [99].

These three systems will be the focus of our binding affinity predictions due to the availability of crystal structures and K_d measurements across different peptides.

4.3.3 H-CNN K_d prediction

We hypothesize that H-CNN-predicted pseudo-energies should correlate with experimental binding affinity since these K_d values are once again related to the free energy of the protein complex in a similar way that $\Delta\Delta G$ and relative fitness were related to the energy for T4 Lysozyme and SARS-CoV-2 RBD in sections 3.2 and 3.3. Specifically, the change in free energy of TCR-pMHC binding is given by $\Delta G = RT \log K_d/C_0$ where T is the temperature, R is the ideal gas constant, and C_0 is a reference concentration that is chosen to be $C_0 = 1$ μM here. Thus we expect the pseudo-energies to correlate with the logarithm of the K_d values. We will use $\log_{10}$ since this only introduces a constant and $\log_{10}$ is used in the literature for its ease of interpretation in terms of concentration.

To compute H-CNN energies, we started from crystal structure templates that featured the same MHC and TCR. In the case of MART-1-like peptides, we also conditioned our

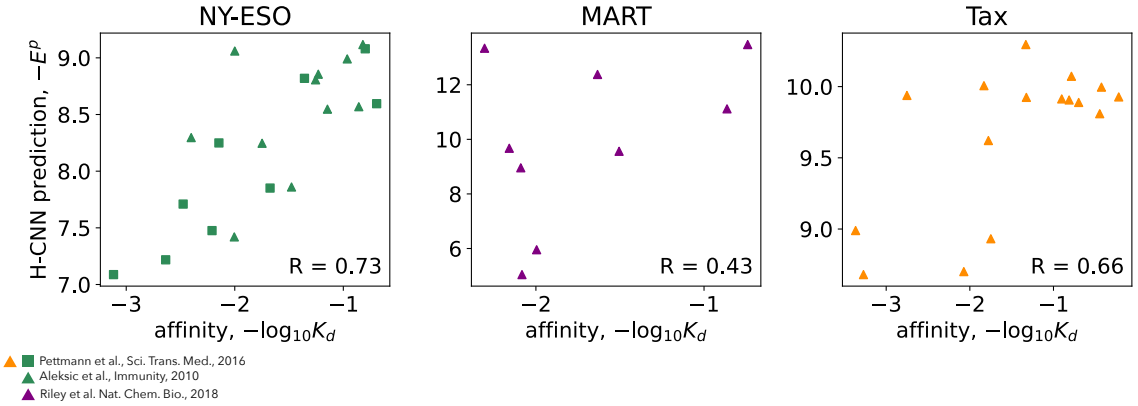


Figure 4.2: **H-CNN energies vs experimentally determined affinities.** H-CNN energies are plotted vs experimentally measured binding affinities for the three antigens NY-ESO-1, MART, and Tax. For NY-ESO, data from [3] and [99] are marked with triangles and squares respectively. Overall, H-CNN energies show good correlation with each system.

structural template to feature a peptide sequence that bore the same sequence motif and structural conformation (see section 4.3.4 for further details). We then mutated the existing peptide that was present in the crystal to the desired peptide via PyRosetta mutation and relaxation. After one relaxation step, we applied H-CNN to every site in the structure and computed energies as defined in equation 3.4. In calculating the network energy, we can consider the entire system or restrict the sum to specific chains in the multimer complex. We found restricting energies to just the peptide gave the best correlations to the experimental K_d. Thus, we define peptide network energy to be

$$E^p(\mathbf{x}) = \frac{1}{\Omega}\sum_{i=1}^{\Omega} E_i^{\sigma_i^p}(\mathbf{x}) \tag{4.2}$$

where σ^p is the peptide sequence and Ω is the length of the peptide sequence.

Restricting our analysis to these peptide energies, the H-CNN computed energies correlate with experimental K_d values ranging from 0.48 to 0.90 for Pearson's R indicating good predictive power over relative binding affinities of different peptides to the same TCR. We show our performance over all datasets in Fig. 4.2.

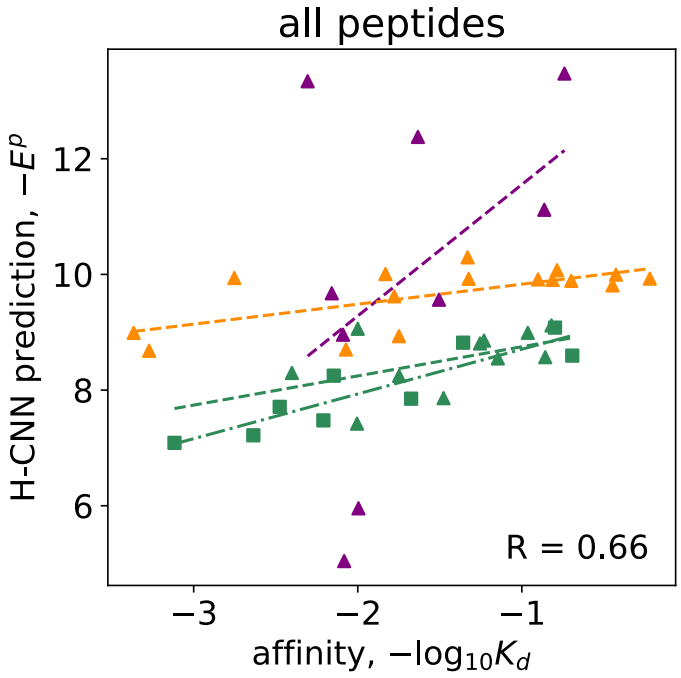


Figure 4.3: **Comparison of linear fits of H-CNN energies vs binding affinities.** Plotting all systems on the same axis reveals the lines of best fit differ from system to system but not between studies of the same system. This difference is evidence that there is a system-dependent scale factor that prevents mapping of H-CNN energies to absolute affinities.

We then investigated whether H-CNN energies can map directly to absolute binding affinities. Comparing linear fits across studies and across systems, we note that H-CNN predictions are comparable across the two studies that measure NY-ESO-1 affinities but not across studies that measure different affinities of different TCRs; see Fig. 4.3. From this analysis, we see that peptide energies normalized by sequence length are not comparable across systems with different peptides and TCRs. We hypothesize that this is due to a system-dependent scale factor that is not captured by H-CNN energies.

4.3.4 H-CNN energies of different MART peptide motifs reflect known structural differences

HLA-A*02-presented MART-like peptides binding to TCR DMF5 have been observed to bind in two distinct modes [104] as visualized in Fig. 4.4A. Specifically, peptides displaying p4-6 DRG sequence motif are shown to undergo a register shift (or a shift in the placement of the peptide in the MHC along length of the pocket) compared to peptide displaying p4-6 GIG. PDB structures 3QDG [18] and 6AM5 [104] feature DMF5 bound to GIG peptides while the crystal structure 6AMU [104] features a DRG peptide. The difference in peptide conformations is quantified by the RMSD of backbone atoms between the structures. The GIG structures have a RMSD to each other of 0.29 Å while they have an RMSD of 1.84 Å and 1.45 Å to the DRG structure.

H-CNN is sensitive to the structural differences in the peptide conformations as demonstrated by its different predictions on each structure. Fig. 4.2 and the calculated Pearson R of 43% reflect using GIG structures for GIG peptides and DRG structures for DRG peptides. If we were to only use GIG or DRG structures, the correlations would be 47% and 0% respectively (Fig. 4.4B,C). Although the correlation is still positive in the case of using GIG structures, we hypothesize that this correlation is spurious since DRG peptides have weaker affinities than GIG peptides in general. This hypothesis is supported by the lack of correlation when using the DRG structure as our initial structure.

This analysis reveals that not only can H-CNN predict the relative binding affinities of different peptides to a given TCR but that its predictions are sensitive to small scale conformational changes in the peptide. Thus, we can conclude that on the limited data for which both crystal structures and experimentally measured binding affinities exist, H-CNN can be used to predict relative affinities of peptides that bind to a given TCR when presented in a given conformation by given MHC.

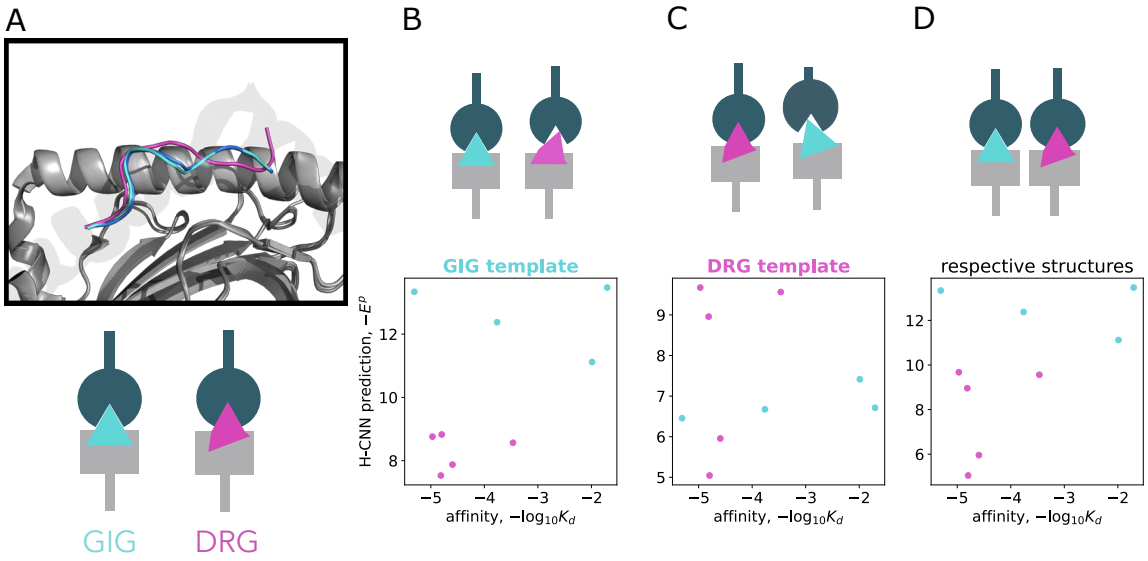


Figure 4.4: **H-CNN predictions conditioned on MART-like peptide conformations reveal sensitivity to structure.** (A) MART peptide conformations are illustrated in a schematic. The proper conformation is necessary to the TCR to recognize each peptide. (B) When using the structures displaying GIG as a template for peptide relaxation, we expect worse recognition by the TCR for sequences baring the DRG motif. This is reflected in H-CNN predictions of DRG sequences. We hypothesize that the correlation recovered in this case is spurious due to the generally worse binding of DRG sequences. (C) Meanwhile, predictions made using the DRG structure should favor DRG sequences. While DRG sequences are not absolutely favored, they are favored relatively to predictions made using the GIG structure. (D) Finally, using the appropriate structure for a given sequence, we see that H-CNN does indeed recover positive correlation with experimentally measured affinities.

4.4 Decoy peptide binder discrimination

4.4.1 The TCR-pMHC docking problem

The problem of docking TCR-pMHC complexes is posed as predicting the co-complexed structures from the TCR and pMHC sequences. The difficulty of predicting these multimeric complexes stands in contrast to folding a single protein, which has been solved [66, 8, 77] in part using sequence homology. Slight sequence variations in CDR regions can produce markedly different structures and therefore binding modes, largely nullifying the utility of sequence homology. The ability to accurately and efficiently score predicted structures produced by TCR-pMHC docking models would be a significant contribution to the field of TCR-pMHC docking. Because H-CNN performed well predicting the relative affinities of TCR-pMHC complexes based on structure alone, it appears to have potential for scoring the viability of TCR-pMHC complexes produced *in silico*.

4.4.2 Decoy dataset

We evaluated H-CNN on a subset of decoy binders generated from an extension of AlphaFold Multimer to specifically fold TCR-pMHC complexes [19]. Specifically, for eight different wild-type pMHC systems, we presented H-CNN with structures of 50 TCRs bound to a target peptide and nine decoy peptides; see Fig. 4.5 for a schematic of the data. Decoy peptides were generated using NetMHCpan-4.1 as described in [19]. We then calculated H-CNN predicted peptide energies as described in equation 4.2. A summary of systems used can be found in Table 4.2.

Overall, we found H-CNN's ability to discriminate target from decoy structures varied from system to system with the strongest dependence on the peptide length. Specifically, H-CNN was able to discriminate targets from decoys with an AUROC of 0.53 and 0.84 for 9-mer and 10-mer peptides respectively. The contrast in H-CNN's performance on this discrimination task with its ability to predict relative binding affinities raises interesting questions regarding H-CNN, the computational prediction of TCR-pMHC structures, and the nature of TCR-pMHC binding. We considered three hypotheses for this discrepancy in performance.

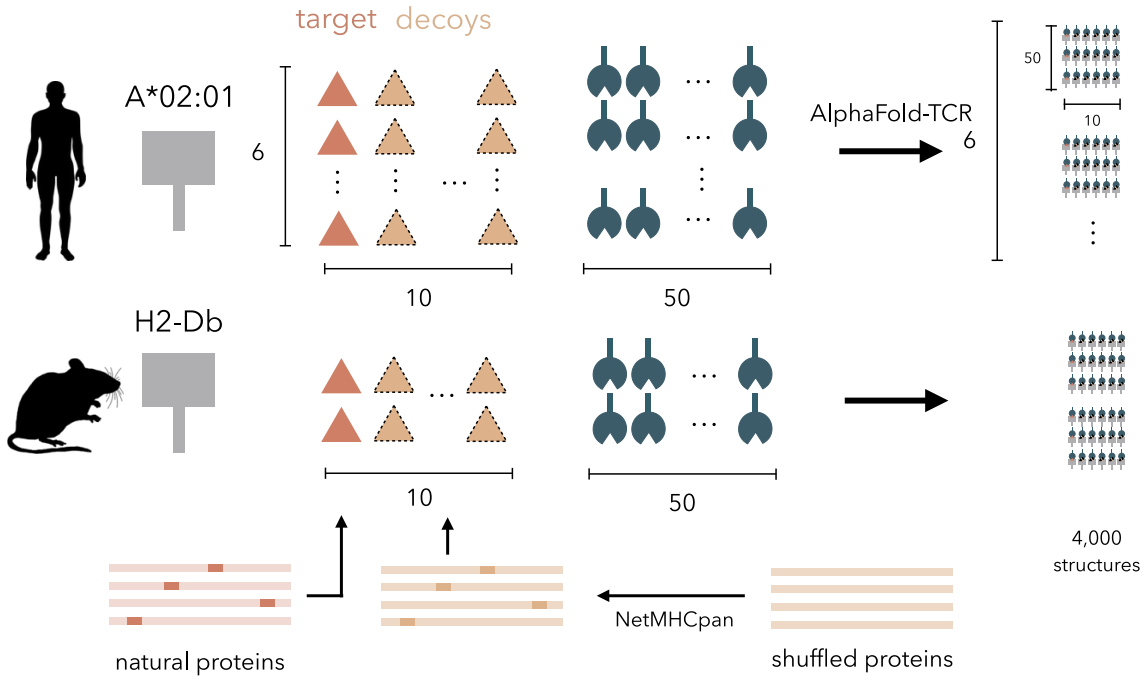


Figure 4.5: **Overview of decoy dataset used to evaluate H-CNN's ability to score docked structures.** Two MHC alleles were studied: the human A*02:01 and murine H2-Db. For each allele, six and two peptides were chosen, each of which had 50 known TCRs binders. In addition to these eight peptides, nine decoy peptides per target were generated from NetMHCpan4.1 selection of possible epitopes from shuffled protein sequences. AlphaFold-TCR [19], an extension of AlphaFold-Multimer, generated docked structures.

Organism	MHC	L	Peptide	Protein	H-CNN AUROC	AlphaFold AUROC
human	HLA-A*02:01	9	GILGFVFTL	Flu M1	0.795	0.822
human	HLA-A*02:01	9	GLCTLVAML	EBV BMLF1	0.293	0.647
human	HLA-A*02:01	9	NLVPMVATV	CMV pp65	0.588	0.578
human	HLA-A*02:01	9	YLQPRTFLL	SARS-CoV-2 Spike	0.666	0.966
human	HLA-A*02:01	10	ELAGIGILTV	human MART-1	0.957	0.960
human	HLA-A*02:01	10	KLVALGINAV	HCV POLG	0.840	0.800
mouse	H2-Db	9	ASNENMETM	Flu NP	0.496	0.750
mouse	H2-Db	10	SSLENFRAYV	Flu PA	0.941	0.947

Table 4.2: Summary of H-CNN performance on decoy discrimination task. For each system, we calculated the AUROC of H-CNN. H-CNN discriminates 10-mer peptides better than 9-mer peptides. AUC for AlphaFold-Multimer-TCR was taken from [19]. H-CNN and AlphaFold-Multimer-TCR discriminate better on 10-mers than on 9-mers although to different extents. Generally, AF-Multimer-TCR outperforms H-CNN.

One hypothesis for reconciling the difference in performance may lie in the system-dependent scale. In correlating H-CNN energies with binding affinities (section 4.3), we observed that H-CNN could rank the relative affinities of peptides with few mutations to the same TCR but could not absolutely predict binding affinity across systems with different TCRs and antigens. Since the decoy peptides here vary by more than a few mutations, it is possible that H-CNN's poor performance is further evidence that H-CNN's energies are not always comparable as both TCR and peptide vary across the set of target and decoy structures for each system. However, this hypothesis would not immediately explain the length dependent variation in performance since all systems display similarly variable peptides.

A second hypothesis explaining the relatively poor performance on 9-mer systems is that the generated decoys actually are mimotopes meaning that they are good binders theoretically but are not necessarily epitopes due to their non-existence in natural protein sequences. Unfortunately such a hypothesis would require experimental testing to answer conclusively.

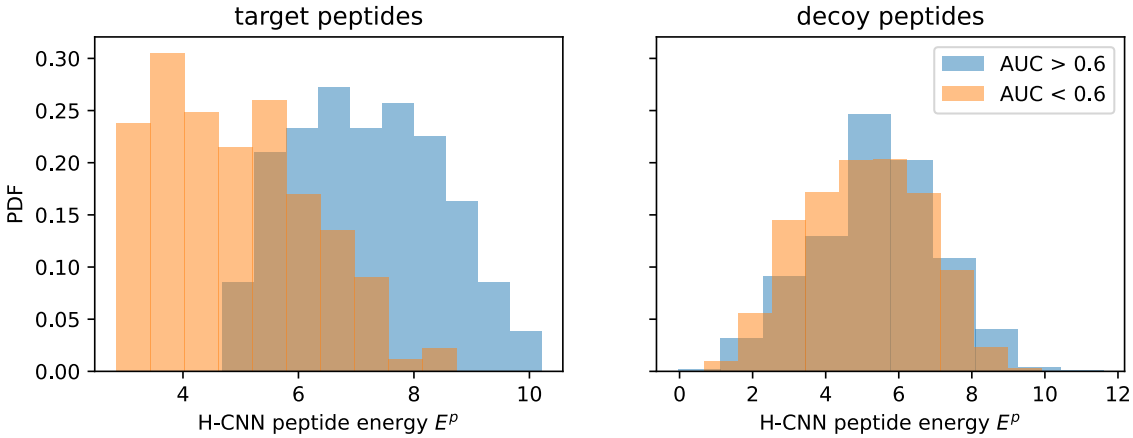


Figure 4.6: **Comparison of peptide energies for target and decoy for systems well and poorly classified by H-CNN.** Distributions for peptide energies are shown for target (left) and decoy (right) structures. Distributions are conditioned on H-CNN's ability to classify targets from decoys as measured by AUROC being above or below 0.6 (blue and orange respectively). Comparison of target peptide energies reveals that the systems H-CNN cannot discriminate have poor peptide energies markedly different from systems H-CNN can discriminate and from crystal structures. Since the distribution of H-CNN's predictions on decoy structures are similar regardless of the system, this analysis suggests that H-CNN's inability to discriminate decoys from targets may be due to the poor quality of the target structures.

Our third hypothesis relates to the structural prediction method used to dock the TCR-pMHC complexes. First, we note that H-CNN's discriminatory power also falls short of AlphaFold-Multimer-TCR's ability to discriminate decoys via its confidence score plDDT with an overall AUC of 0.6 vs 0.8 (see Table 4.2 for a system-by-system comparison). If AlphaFold's confidence metrics, which have also been shown to be correlated with stabilities [110], can discriminate decoys from binders, then what is preventing H-CNN from discriminating these peptides? To answer this question, we turn our focus to a system-by-system analysis of H-CNN's errors. Specifically, we note that the three worst classified systems (GLCTLVAML/EBV BMLF1, ASNENMETM/Flu NP, NLVPMVATV/CMV pp65) yield a distribution of H-CNN target peptide energies that is markedly lower than the distribution peptide energies on crystal structures and other AlphaFold-TCR target structures (see Fig. 4.6A).

While the deviations of these energies alone do not prevent H-CNN from classifying well, it is observed that H-CNN predicts target peptides in these systems to have lower energies than general decoy peptides (see Fig. 4.6B). This relative prediction is the source of H-CNN's error. Interestingly, AlphaFold's confidence metrics appear to similarly indicate that the target peptides yield relatively worse structures than the target peptides of other systems, however AlphaFold-TCR can still recognize decoys as giving worse structures. Taken together, in the cases where H-CNN fails to discriminate, both AlphaFold-TCR and H-CNN recognize the structures as relatively worse than other systems. However, only H-CNN fails to recognize the decoys in these systems to be even worse. While this result may reflect AlphaFold-TCR's superior ability to reason about structures, we also hypothesize that it may reflect the advantage of using sequence information especially in a case where the decoy peptides are not taken explicitly from natural sequences.

4.5 Generation of binding peptide sequences

The specificity of a TCR is defined by the peptide sequences to which it binds. When studying TCR specificity, it is natural to ask if this specificity can be computationally characterized. Specifically, given a TCR, can one produce the peptide sequences to which it would bind? Since there are $20^9 \approx 10^{12}$ possible peptides of just length 9, scoring all

possible peptides is intractable and sampling is difficult. In this section, we propose a method to generate peptides based on the energy landscape as defined by H-CNN in hopes of determining the specificity of TCRs.

4.5.1 Using H-CNN energy landscape to sample sequence space

Assuming all possible peptides were presented to a specific TCR, we would expect the TCR to be bound to a given peptide with sequence σ with probability dependent on the free energy of binding ΔG. The simplest model would be an equilibrium one where the probability is given by the Boltzmann distribution

$$P(\sigma) = \frac{e^{-\Delta G^{\sigma}}}{\sum_{\sigma'} e^{-\Delta G^{\sigma'}}} \tag{4.3}$$

where σ' is summed over all possible sequences. In section 4.3, we demonstrated that H-CNN generally has decent predictive power towards TCR-pMHC binding affinity when presented with an accurate structure. In particular, the energies predicted by the model $E^{\sigma^p}(x)$ are correlated with physical binding energies $-\log K_d \propto \Delta G$ and thus for a given sequence we can approximate the numerator of the probability. However, even if we knew the exact ΔG of all sequences, we could not compute the denominator due to the size of sequence space. Thus, H-CNN gives access to an approximate unnormalized probability distribution over sequence space.

The problem of sampling from unnormalized probability distributions is well studied in statistics and machine learning. For example, the Metropolis-Hastings algorithm [87, 52] is a Markov chain Monte Carlo (MCMC) method that allows one to sample from an unnormalized probability distribution. One can also use Langevin dynamics to sample from unnormalized probability distributions [151, 120].[4]

Since H-CNN is a classification model that predicts the probability of amino acids at a specific site in proteins, we can only directly use it to sample from sequence space given a structure. Sepcifically, we would like to sample

$$\sigma \sim P_\theta(\sigma|\mathbf{x}) = \frac{e^{-E^{\sigma}(\mathbf{x})}(\mathbf{x})}{\sum_{\sigma'} e^{-E^{\sigma'}(\mathbf{x})}} \tag{4.4}$$

[4]Diffusion models have recently shown enormous success using Langevin dynamics but rather than learning the unnormalized probability distribution, they learn the gradient directly [60, 122, 80].

where $E^{\sigma}(x) = \sum_i E^{\sigma_i}(x)$ is the sum of H-CNN predicted energies at each site i, $\mathbf{x}$ is the structure of the complex, and σ' is summed over all possible sequences. As mentioned above, the denominator is intractable to compute and thus our classifier can only give a unnormalized probability distribution over sequence space. Furthermore, since we are using a crystal structure as a starting structure for our initial evaluation of energies, we would like to incorporate the epistatic effects that mutations in the peptide sequence may have. Thus, we would actually like to sample from the joint distribution $P(\sigma, \mathbf{x})$ with the assumption that the structures we will sample will be constrained to be somewhat similar to the initial structure.

With these constraints in mind, we propose a simulated annealing sampling procedure where we will use the energies defined by H-CNN to explore sequence space and PyRosetta structural relaxations to sample structure space.

4.5.2 Simulated annealing

Simulated annealing is an stochastic approximate global optimization scheme that is based on the physical process of annealing. In particular, given a system's state space $\mathcal{S}$ and an energy function $E(s)$ over the states $s \in \mathcal{S}$, simulated annealing generally searches for the global minimum of $E(s)$ by starting at a random state $s^{(0)}$ and iteratively moving to a new state s' with a probability defined by the difference in energies $\Delta E = E(s') - E(s)$ and a temperature T. The temperature T defines how likely transitioning to an unfavorable state is and as it is slowly decreased, transitions to unfavorable states become less and less likely with the system being fixed in a local optimum when the temperature reaches zero.

Simulated annealing algorithms are simple to implement and are often favored in problems that feature discrete state spaces, however they do often suffer from slow convergence rates. As proof of principle, simulated annealing serves as a good starting point for our generative model.

Our simulated annealing algorithm is designed to generate peptide sequences that are likely to bind to a given TCR-pMHC complex. To generate sequences, we first initialize the structure of the complex to be the crystal structure $\mathbf{x}^{(0)}$, the peptide sequence to be a

random sequence $\sigma^{(0)}$ with each amino acid equally likely at each site, and the temperature to be $T^{(0)}$. This peptide sequence is input into the structure *in silico* via PyRosetta's packing functionality and PyRosetta's `Relax` protocol is then used to optimize the structure for this new sequence [22]. Specifically, we allow for movement of all atoms in the peptide, while we only allow side chain torsional degrees of freedom in the MHC and TCR. After a new structure is produced, we then evaluate H-CNN on all sites in the peptide and sample a sequence $\sigma^{(t+1)}$ where each amino acid $\sigma_i^{(t+1)}$ is drawn from a multinomial distribution at each site given by the H-CNN-defined energies. This new peptide is packed and the structure is relaxed yielding a new structure $\mathbf{x}$. We then evaluate H-CNN on the new structure at both peptide and pocket sites (i.e., all sites within 10 Å of the peptide alpha carbons). If the new structure is more favorable than the previous structure $\mathbf{x}^{(t)}$ according to H-CNN then the new structure is automatically accepted to be $\mathbf{x}^{(t+1)}$. If the new structure is not more favorable then it is only accepted with probability $e^{-\Delta E/T^{(t)}}$ where ΔE is the difference in energies between the new and old structures. The temperature $T^{(t)}$ is then adjusted according to a schedule and the process is repeated until the temperature reaches zero. The full algorithm is given in Algorithm 1 and temperature schedules are discussed in the following subsection.

4.5.3 Cooling schedules

In general, the optimal cooling schedule for a simulated annealing algorithm depends on the problem itself as well as on the computational resources available. Here we test two cooling schedules–one predetermined schedule and one that adjusts to the annealing history itself.

Multiplicative cooling The first schedule we tested is a multiplicative cooling schedule where the temperature decays from some T_0 to T_f under a power law based on some parameter a which is specified prior to annealing. Specifically,

$$f_T(k; T_0, T_f, a) = (T_0 - T_f)a^k + T_f. \tag{4.5}$$

Algorithm 1 Simulated annealing algorithm for generation of peptide sequences

$T_{\text{peptide}} = T^{(0)}_{\text{peptide}}$
$T_{\text{pocket}} = T^{(0)}_{\text{pocket}}$
$\boldsymbol{\sigma}^{(0)} =$ random sequence
Mutate structure via rosetta
$\mathbf{x}^{(0)} =$ FastRelax structure via rosetta
$E_{\text{pocket}} = \sum_{i \in \mathcal{R}_{\text{pocket}}} E_{\alpha_i}(\mathbf{x}_i)$
Calculate $\mathbf{E}(\mathbf{x}_i)$ via H-CNN for each $i \in \mathcal{R}_{\text{pep}}$
for $k = 1, 2, \ldots$ **do**
    for $j = 1, 2, \ldots, L$ **do**
        $\sigma_i \sim$ argmax Mult$(1, \text{softmax}(\beta_{\text{peptide}}\mathbf{E}(\mathbf{x}_\text{i})))$
    end for
    Mutate peptide to $\boldsymbol{\sigma}$
    $\mathbf{x} =$ FastRelax
    Calculate new $\mathbf{E}(\mathbf{x}_i)$ for each $i \in \mathcal{R}_{\text{pep}}$
    $\mathbf{E}_{\text{pocket}} = \sum_{i \in \mathcal{R}_{\text{pocket}}} \mathbf{E}(\mathbf{x}_i)$
    $p \sim \text{Uniform}(0, 1)$
    if $p < \exp(-\Delta E_{\text{pocket}} \beta_{\text{pocket}})$ **then**
        $E^{(k)}_{\text{pocket}} = E_{\text{pocket}}$
        $\boldsymbol{\sigma}^{(k)} = \boldsymbol{\sigma}$
    else
        $E^{(k)}_{\text{pocket}} = E^{(k-1)}_{\text{pocket}}$
        $\mathbf{x}^{(k)} = \mathbf{x}^{(k-1)}$
        $\boldsymbol{\sigma}^{(k)} = \boldsymbol{\sigma}^{(k-1)}$
    end if
    $T^{(k+1)}_{\text{peptide}} = f_t(T^{(k)}_{\text{peptide}}, A^{(k)}, k)$
    $T^{(k+1)}_{\text{pocket}} = f_t(T^{(k)}_{\text{pocket}}, A^{(k)}, k)$
end for

Thermodynamic schedule To avoid tuning the temperatures to the energy scale of the complex, we also implement Thermodynamic Simulated Annealing (TSA) as prescribed by [32]. The cooling schedule for TSA is a function of how much the energy has changed in the accepted trajectory of the system, the increase in entropy due to proposed unfavorable states, and a parameter k_A which tunes the speed of cooling. Specifically the schedule is given by

$$f_T(\mathbf{T}, \mathbf{A}, \mathbf{E}) = \begin{cases} T_0 & \text{if } \sum_{i \in M_{\text{acc}}} \Delta E_i \leq 0 \text{ or } \sum_{i \in M^+} \Delta \frac{E_i}{T_i} \leq 0 \\ -k_A \frac{\sum_{i \in M_{\text{acc}}} \Delta E_i}{\sum_{i \in M^+} \Delta \frac{E_i}{T_i}} & \text{otherwise} \end{cases} \tag{4.6}$$

where E_i is the H-CNN defined peptide energy at step i, M_{acc} are proposed states that were accepted, and M^+ are proposed states that had an increased energy compared to the current state at the time of proposal.

4.5.4 H-CNN-generated sequences share similar qualitative features to natural sequences

Using the simulated annealing scheme described above, sequences were generated for two systems: peptides that would bind to DMF5 when presented by HLA-A*02:01 and peptides that would bind to the ankylosing spondylitis (AS) associated TCRs when presented by HLA-B*27:05. We chose these systems because high-throughput peptide scans have been carried out for each system [44, 161] yielding datasets for validation of the generative method. For each system, 369 sequences were generated using both the multiplicative cooling schedule and the thermodynamic schedule. We then compared the sequences generated by the two cooling schedules to each other and to the experimentally produced sequences. Convergence to local optima was assessed for the multiplicative cooling schedule annealing runs by measuring the number of mutations made in the low temperature limit. In the thermodynamic cooling runs, convergence was similarly assessed however the low temperature limit reached was also assessed since these schedules do not necessarily trend to $T = 0$. In both cases, annealing runs converged to local optima. Sequence logos that display the frequency of amino acids at each site are shown for both the generated peptides and the natural peptides in Fig. 4.8.

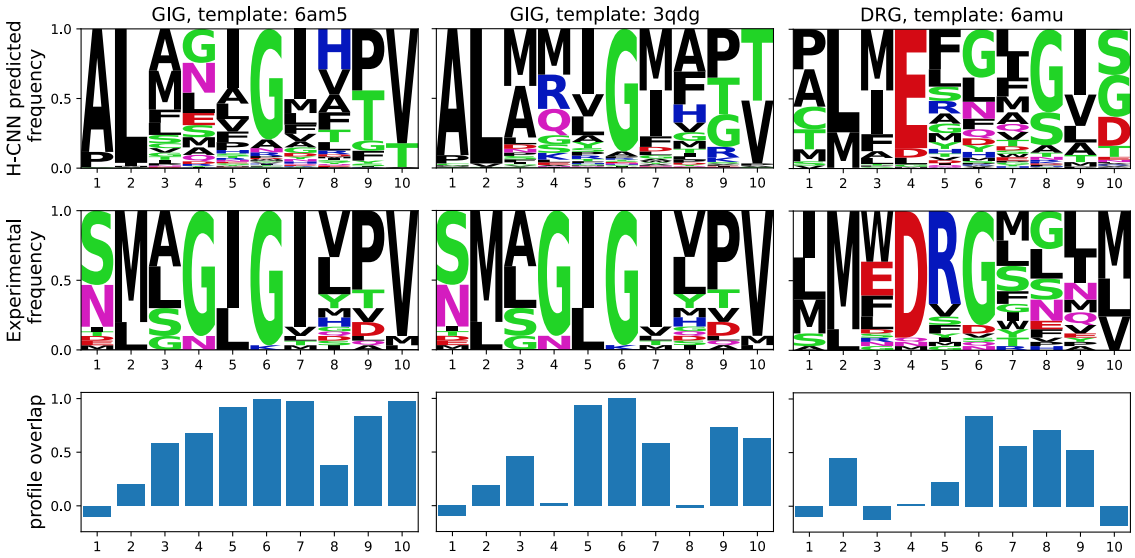


Figure 4.7: **Experimental and annealed 10-mer sequence logos for TCR DMF5.** Sequence logos featuring the frequency of each amino acid at each site in the peptide give by the annealed sequence (top row) and the experimentally determined sequences (middle row). The size of each letter represents the frequency with which that amino acid is found at that site. Amino acid abbreviations are colored according to their chemical properties. Green are polar, purple are neutral, blue are basic, red are positive, and black are hydrophobic amino acids. To quantify the agreement of the predictions with the experimental results, the profile overlap at each site is shown in the bottom row.

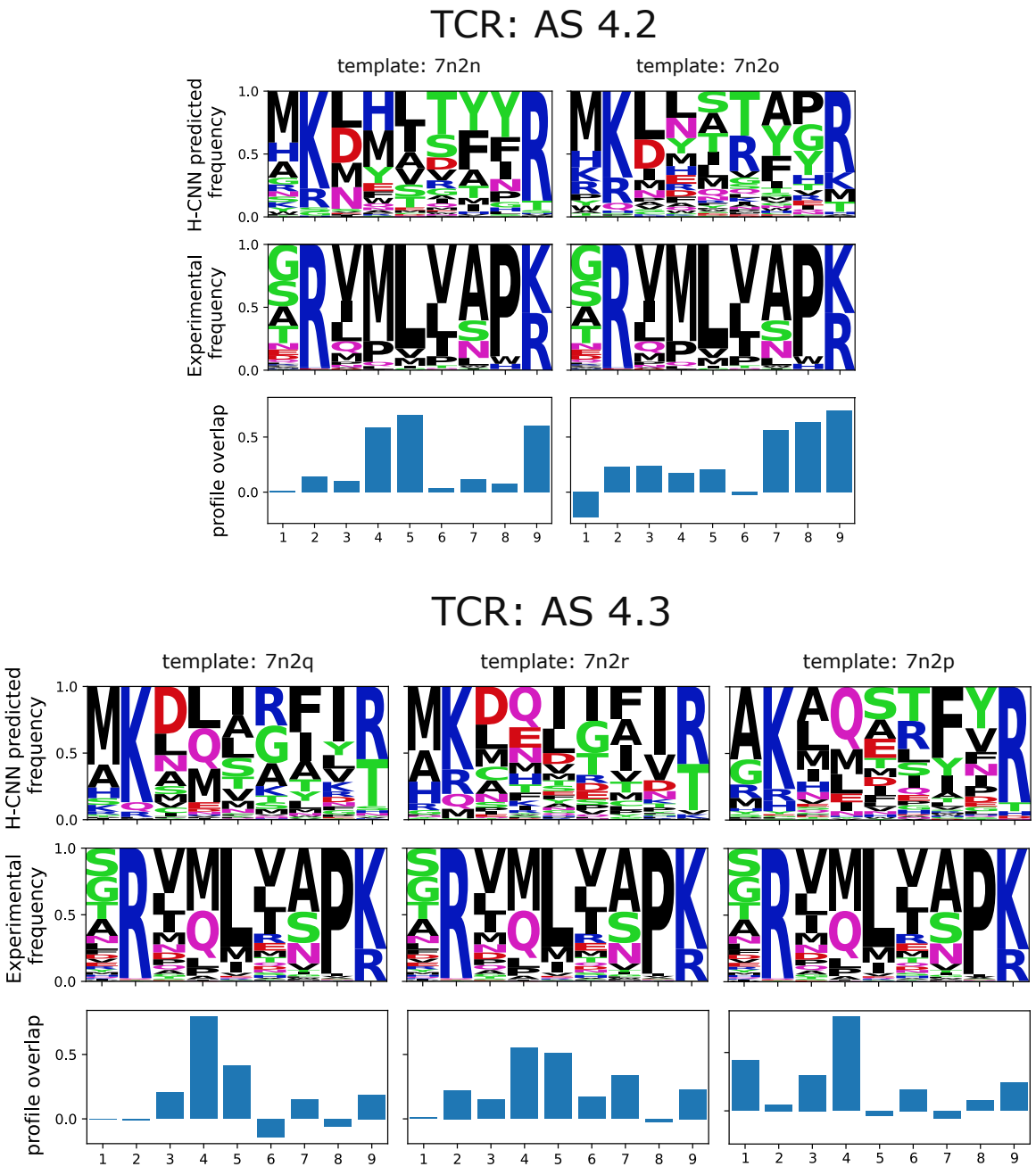


Figure 4.8: **Experimental and annealed 9-mers for TCRs AS4.2 and AS4.3.**

The annealed peptides generally capture qualitative features of the experimentally determined sequences. Using GIG-conformation structures, H-CNN generates sequences displaying a p4-6 GIG motif (see Fig. 4.7). Furthermore, H-CNN sequences at sites p3 and p7 match the experimental sequences quite well recapitulating preferences for hydrophohic and polar amino acids. Finally, H-CNN also predicts the N-terminal well recognizing the importance of proline and valine at sites p9 and p10.

Similar analysis can be carried out for DRG peptides; however the agreement with DRG peptides is noticeably worse than for the GIG conformations. This may be due to the fact that DRG peptides are weaker binders or due to the fact that they undergo a register shift upon binding making the preferences of amino acids dependent on dynamics not accounted for in this generative scheme.

Discrepancies like these become more noticeable as we extend our analysis to AS-reactive TCRs 4.8. While the annealed peptide logos still have a number of qualitative features that agree with the experimental peptides, it is clear that there are dramatic differences between the two sets. For example, in AS/B*27:05 bound peptides, p2 and p9 are favored to have positive residues experimentally and computationally; however H-CNN confuses the use of arginine and lysine at each site favoring one where the opposite is favored in reality. Meanwhile the relatively conserved methionine and leucine at sites p4 and p5 are not recovered as dramatically in the annealed peptides.[5]

4.5.5 Reconciling H-CNN's ability to predict binding affinity with its inability to recapitulate experimental sequences

Considering the significant correlations of H-CNN predicted energies with binding affinities, we can wonder why H-CNN generated peptides are sometimes quantitatively distinct from experimentally sampled peptides. We consider three possible reasons for such a discrepancy: (i) inaccuracy of the model, (ii) unwarranted assumptions, or (iii) the underlying nature of TCR-pMHC interactions.

[5]Also at site p1, methionine is generally favored despited little use of methionine in the experimental peptides. We believe this is due to the strong bias of H-CNN seeing methionine at the N-terminus of complete proteins as opposed to peptide fragments.

First, let us consider possible sources of inaccuracy with the model. H-CNN was only trained on crystal structures and therefore it may not necessarily be accurate in processing non-crystal structures. H-CNN did demonstrate reduced but similar predictive power of mutational effects on stability when using Rosetta-relaxed structures (Fig. 3.4). An H-CNN model trained on Rosetta-relaxed structures or even nuclear magnetic resonance (NMR) spectroscopy structure may yield a model more suitable for this generation task.

It may be instead that our generative scheme has restrictive and unwarranted assumptions. For example, even though side chain flexibility in the TCR and MHC were sufficient for predicting K_d, these restrictions may be limiting in this generative scheme. Just as we observed that starting from a GIG-MART structure never yielded DRG peptide and vice versa, it may be that more structural flexibility is needed in order to best reflect the experimental sampling.

Finally, it may be that the nature of peptide selection is simply not captured by our model. In comparing H-CNN energies with sequence counts, we must recognize that the counts of sequences present in a yeast display assay are not necessarily correlated with binding affinities. For example, comparing $-\log(K_d)$ measured for ten peptides in [104] with the logarithm of the counts of the sequences after three rounds of selection in a yeast phage display [44], we find the Pearson correlation of the two quantities to be $R = 0.17$ meaning that not even the measured K_d values are correlated with experimental counts of sequences. The ability for multiple pMHCs to be displayed on the surface of one cell and the use of TCR tetramers both imply that avidity is the more important factor in T cell activation than binding affinity. A model that hopes to accurately predict T cell specificity will most likely need to account for these cellular scale factors.

In conclusion, this study into TCR specificity prediction has revealed both the abilities and shortcomings of applying a rotationally equivariant structure-based model in a zero-shot setting to the problem of TCR specificity. We demonstrated that H-CNN has significant power at predicting binding affinities of peptides to TCRs when presented with an accurate structure. We have also shown that this model can be used to discriminate binding peptides from decoy TCRs in some cases. Finally, we have shown that the task of predicting counts of sequences from high-throughput TCR-pMHC screening experiments lies outside of H-CNN's

current capabilities. While H-CNN can certainly be trained more thoroughly, the limited data we've analyzed points out a larger problem for investigating TCR specificity with molecule-based machine learning models: binding affinity is just one step of a larger physical process. We hypothesize that machine learning models will have to move beyond binding affinity towards avidity to accurately predict T cell specificity.

Chapter 5

CONCLUSION AND OUTLOOK

The ability to predict the quantitative function of a protein is extremely important for both protein science and engineering. Historically, this prediction problem has been challenging due to the complexity of protein physics and chemistry. First-principles models of atomic interactions are sufficiently accurate but are too computationally expensive to be practical. As computational power continues to increase, data-driven methods are an attractive alternative due to their relatively cheap cost. However, these methods have still not yet solved function prediction. Sequence-based methods are challenged by the complex and indirect relationship between primary structure and function and struggle with generalization due to biases in sequence datasets. Structure-based models are asked to learn the more direct map from structure to function, yet still these models have struggled to achieve high accuracy in prediction in part due to the high dimensionality of 3D structures and the relatively low abundance of structural data.

The emerging field of geometric deep learning offers a new class of data-driven structure-based methods that address the curse of dimensionality that plagues traditional structure-based ML models. Specifically, these methods utilize the physical principle of symmetry to reduce the dimensionality of the data and of the model search space. Incorporation of symmetry in these models also endows these ML models with many of the attractive properties of traditional physics models such as interpretability and generalizability. In this thesis, I have attempted to build such a model to ultimately address the task of protein function prediction.

In chapter 2, I formulated H-CNN, a minimal rotationally equivariant ML model for processing all-atom representations of protein structures. Trained to predict amino acid identities based on atomic surroundings, H-CNN exhibits state-of-the-art predictive power in its task and superior efficiency both in terms of training time and the number of parameters

used. In this sense, H-CNN is evidence that geometric deep learning models offer efficiency without sacrificing performance. Furthermore, the analysis presented in chapter 2 shows H-CNN is useful beyond its predictive power and efficiency. In particular, I have shown that H-CNN can extract information from atomic environments that is related to the physical and chemical properties of the amino acids. The structure of H-CNN predictions also reflect the evolutionary use of amino acids. Coupled with the theoretical generalizability and interpretability of equivariant models, these results suggest that H-CNN may have learned an effective potential for amino acids.

In chapter 3, I developed an interpretation of H-CNN as an effective free energy potential for amino acids in atomic environments and tested the validity of such an interpretation. Through this investigation, I demonstrated that not only are H-CNN pseudo-energies consistent with an effective potential in their physical response to perturbations of structures, but they correlate with experimentally measured free energies of stability and binding. This correlation is notable since it means H-CNN has predictive power over quantitative protein energies in a zero-shot setting. That is to say that H-CNN can predict energies without ever having been explicitly trained to predict them. This ability is evidence of interpretability and generalizability of equivariant models that make them attractive methods.

Recognizing the utility of a generalizable effective potential for proteins, I then turned in chapter 4 to apply H-CNN to investigate the notoriously challenging problem of predicting TCR specificity. Consistent with H-CNN ability to predict binding in the CoV2-RBD-ACE2 complex, H-CNN also has good predictive power over relative binding affinities of different peptides to a given TCR—a notable result not just because of the zero-shot nature of the prediction, but also because of the complexity of TCR-pMHC binding. Never having been trained explicitly on TCR-pMHC complexes, H-CNN is able to rank the relatively weak binding affinities of many peptides in a larger protein complex of TCR, peptide, and MHC.

I then demonstrated that H-CNN has potential to guide docking methods for TCR-pMHC structures. H-CNN's varied ability to score docked structures raises interesting questions about TCR-pMHC structure prediction methods and the nature of decoy binders used by these methods: Is sequence or template information significantly advantageous in scoring docked structures? How does the effective potential that H-CNN has learned compare to

effective potentials that appear to exist in other ML models such as AlphaFold [110]? And how prevalent are mimotopes in sets of decoy peptides such as the one analyzed here? This analysis has also raised a number of interesting ideas for improving and extending H-CNN which I will discuss momentarily.

In the last part of chapter 4, I continued the energetic interpretation of H-CNN by asking, if H-CNN can approximate free energy of binding for different peptides, can we sample from this free energy landscape to determine the specificity of a given TCR? I demonstrated that H-CNN can determine qualitative and to some extent quantitative features of the distribution of peptides to which a specific TCR would bind. However, discrepancies in predictions revealed that affinity may not be the ultimate quantity of interest in predicting TCR specificity. Instead, avidity is a more reasonable quantity underlying TCR specificity. While avidity, depending on molecular interactions and cellular geometry, expression, co-receptors, and signaling molecules, is multiscale in nature and therefore out of the scope of H-CNN, the affinity predictions from H-CNN can certainly contribute to a larger multiscale model of TCR specificity.

Beyond, the work presented here the development and testing of H-CNN contributes to the larger field of geometric deep learning applied to protein science. ML of proteins for scientific and engineering purposes is a rapidly growing field. At the time of writing, this model is one of hundreds and at the time of reading, this model may already be one of thousands. In this sea of models, the advantage of geometric deep learning is evident. Almost all state-of-the-art models that involve protein structure incorporate the principles of geometric deep learning to some extent. While this general trend is clear, it is less obvious which model architecture, if any, is optimal for studying proteins at large. As the field searches for the best models and design principles, it important to explore a diverse set of options as ultimately it is unlikely that one model will be absolutely superior in all tasks. Much more likely is the case where some models are superior at solving specific problems. In this context, I believe this thesis contributes towards this exploration of design space.

H-CNN stands out from other protein-specific ML architectures in numerous ways. First, it is one of the few fully Fourier treatments of proteins. This design choice allows for efficient representations of all-atom structures which may be advantageous in certain cases. For

example, while it is clear that backbone structures are often sufficient for many protein engineering tasks, these models are non-amino acid biomolecules such as ligands, nucleotides, metallic ions, and post-translationally modified amino acids. Even though the formulation of H-CNN presented here does not permit these molecules either, the equivariant encoding can easily be extended accommodate them or even continuous densities if one desired.

The flavor of equivariant processing of H-CNN is also unique among rotationally equivariant architectures. Most models that incorporate rotational symmetry just use vector quantities ($L_{\max} = 1$) with some extending to $L_{\max} = 3$ [93]. The use of $L_{\max} = 5$ in H-CNN (and sometimes higher in development) could allow for rich geometry dependent representations of protein structures to be learned for use in tasks beyond quantitative function prediction. For example, invariant features could be used to predict qualitative function. Vector features could be used to determine side chain geometries in rotamer packing tasks. Or higher order features could serve as a useful representation for determining complementarity of protein surfaces aiding docking methods.

The ability to learn efficient invariant representations also is attractive for its somewhat orthogonal perspective on representing residues in proteins. It would be interesting to quantify how much shared versus independent information H-CNN representations at a site in a protein have when compared to representations learned from large protein language models. Ultimately, engineering efforts may want to leverage both sets of representations simultaneously for maximal performance.

There are also a number of extensions to the training of H-CNN that could improve its performance. A contrastive loss objective would allow H-CNN to approximate joint distributions of amino acid and structure as opposed to conditional distributions of amino acid given structure. Such objectives have been shown to have success to machine learning models at large [48] and on protein-specific models [34]. Developing H-CNN in this direction would likely improve its performance in both generative and scoring tasks.

Finally, H-CNN could present an attractive baseline for performing all atom diffusion in protein structures. Diffusion models formulated as gradients of energy functions [7] have been shown to be effective models for learning denoising diffusion probabilistic models. H-CNN's demonstration as an effective potential could serve as good pretrained model for further

development in diffusion models.

BIBLIOGRAPHY

[1] Wajid Arshad Abbasi, Adiba Yaseen, Fahad Ul Hassan, Saiqa Andleeb, and Fayyaz Ul Amir Afsar Minhas. ISLAND: in-silico proteins binding affinity prediction using sequence information. *BioData Mining*, 13(1):20, November 2020.

[2] Rhys M Adams, Thierry Mora, Aleksandra M Walczak, and Justin B Kinney. Measuring the sequence-affinity landscape of antibodies with massively parallel titration curves. *eLife*, 5:e23156, December 2016.

[3] Milos Aleksic, Omer Dushek, Hao Zhang, Eugene Shenderov, Ji-Li Chen, Vincenzo Cerundolo, Daniel Coombs, and P. Anton van der Merwe. Dependence of T Cell Antigen Recognition on T Cell Receptor-Peptide MHC Confinement Time. *Immunity*, 32(2):163–174, February 2010.

[4] Alaa Alnefaie, Sarah Albogami, Yousif Asiri, Tanveer Ahmad, Saqer S. Alotaibi, Mohammad M. Al-Sanea, and Hisham Althobaiti. Chimeric Antigen Receptor T-Cells: An Overview of Concepts, Applications, Limitations, and Proposed Solutions. *Frontiers in Bioengineering and Biotechnology*, 10:797440, 2022.

[5] Mohammed AlQuraishi. ProteinNet: a standardized data set for machine learning of protein structure. *BMC Bioinformatics*, 20(1):311, June 2019.

[6] D. Eric Anderson, James H. Hurley, Hale Nicholson, Walter A. Baase, and Brian W. Matthews. Hydrophobic core repacking and aromatic-aromatic interaction in the thermostable mutant of T4 lysozyme ser 117 → phe. *Protein Science*, 2(8):1285–1290, August 1993.

[7] Marloes Arts, Victor Garcia Satorras, Chin-Wei Huang, Daniel Zuegner, Marco Federici, Cecilia Clementi, Frank Noé, Robert Pinsler, and Rianne van den Berg. Two for One: Diffusion Models and Force Fields for Coarse-Grained Molecular Dynamics, February 2023. arXiv:2302.00600 [cs].

[8] Minkyung Baek, Frank DiMaio, Ivan Anishchenko, Justas Dauparas, Sergey Ovchinnikov, Gyu Rie Lee, Jue Wang, Qian Cong, Lisa N. Kinch, R. Dustin Schaeffer, Claudia Millán, Hahnbeom Park, Carson Adams, Caleb R. Glassman, Andy DeGiovanni, Jose H. Pereira, Andria V. Rodrigues, Alberdina A. van Dijk, Ana C. Ebrecht, Diederik J. Opperman, Theo Sagmeister, Christoph Buhlheller, Tea Pavkov-Keller, Manoj K. Rathinaswamy, Udit Dalwadi, Calvin K. Yip, John E. Burke, K. Christopher Garcia, Nick V. Grishin, Paul D. Adams, Randy J. Read, and David Baker. Accurate

prediction of protein structures and interactions using a three-track neural network. *Science*, 373(6557):871–876, August 2021. Publisher: American Association for the Advancement of Science.

[9] Christopher M. Baker and Guy H. Grant. Role of aromatic amino acids in protein–nucleic acid recognition. *Biopolymers*, 85(5-6):456–470, 2007. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/bip.20682.

[10] Simon Batzner, Albert Musaelian, Lixin Sun, Mario Geiger, Jonathan P. Mailoa, Mordechai Kornbluth, Nicola Molinari, Tess E. Smidt, and Boris Kozinsky. E(3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials. *Nature Communications*, 13:2453, May 2022.

[11] Susanta K. Behura and David W. Severson. Codon usage bias: causative factors, quantification methods and genome-wide patterns: with emphasis on insect genomes. *Biological Reviews*, 88(1):49–61, 2013. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1469-185X.2012.00242.x.

[12] Helen M. Berman, John Westbrook, Zukang Feng, Gary Gilliland, T. N. Bhat, Helge Weissig, Ilya N. Shindyalov, and Philip E. Bourne. The Protein Data Bank. *Nucleic Acids Research*, 28(1):235–242, January 2000.

[13] V Bhardwaj, V Kumar, H M Geysen, and E E Sercarz. Degenerate recognition of a dissimilar antigenic peptide by myelin basic protein-reactive T cells. Implications for thymic education and autoimmunity. *The Journal of Immunology*, 151(9):5000–5010, November 1993.

[14] Gabriela Bitencourt-Ferreira and Walter Filgueira de Azevedo. Machine Learning to Predict Binding Affinity. *Methods in Molecular Biology (Clifton, N.J.)*, 2053:251–273, 2019.

[15] Pamela J. Bjorkman. MHC Restriction in Three Dimensions: A View of T Cell Receptor/Ligand Interactions. *Cell*, 89(2):167–170, April 1997. Publisher: Elsevier.

[16] A. A. Bogan and K. S. Thorn. Anatomy of hot spots in protein interfaces. *Journal of Molecular Biology*, 280(1):1–9, July 1998.

[17] Wouter Boomsma and Jes Frellsen. Spherical convolutions and their application in molecular modelling. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[18] Oleg Y. Borbulevych, Sujatha M. Santhanagopolan, Moushumi Hossain, and Brian M. Baker. TCRs Used in Cancer Gene Therapy Cross-React with MART-1/Melan-A Tumor Antigens via Distinct Mechanisms. *The Journal of Immunology*, 187(5):2453–2463, September 2011.

[19] Philip Bradley. Structure-based prediction of T cell receptor:peptide-MHC interactions. *eLife*, 12:e82813, January 2023. Publisher: eLife Sciences Publications, Ltd.

[20] Michael M. Bronstein, Joan Bruna, Taco Cohen, and Petar Veličković. Geometric Deep Learning: Grids, Groups, Graphs, Geodesics, and Gauges, May 2021. Number: arXiv:2104.13478 arXiv:2104.13478 [cs, stat].

[21] Anna M. Bulek, David K. Cole, Ania Skowera, Garry Dolton, Stephanie Gras, Florian Madura, Anna Fuller, John J. Miles, Emma Gostick, David A. Price, Jan W. Drijfhout, Robin R. Knight, Guo C. Huang, Nikolai Lissin, Peter E. Molloy, Linda Wooldridge, Bent K. Jakobsen, Jamie Rossjohn, Mark Peakman, Pierre J. Rizkallah, and Andrew K. Sewell. Structural basis for the killing of human beta cells by CD8(+) T cells in type 1 diabetes. *Nature Immunology*, 13(3):283–289, January 2012.

[22] Sidhartha Chaudhury, Sergey Lyskov, and Jeffrey J. Gray. PyRosetta: a script-based interface for implementing molecular modeling algorithms using Rosetta. *Bioinformatics*, 26(5):689–691, March 2010.

[23] Ji-Li Chen, Guillaume Stewart-Jones, Giovanna Bossi, Nikolai M. Lissin, Linda Wooldridge, Ed Man Lik Choi, Gerhard Held, P. Rod Dunbar, Robert M. Esnouf, Malkit Sami, Jonathan M. Boulter, Pierre Rizkallah, Christoph Renner, Andrew Sewell, P. Anton van der Merwe, Bent K. Jakobsen, Gillian Griffiths, E. Yvonne Jones, and Vincenzo Cerundolo. Structural and kinetic basis for heightened immunogenicity of T cell vaccines. *The Journal of Experimental Medicine*, 201(8):1243–1255, April 2005.

[24] Muhao Chen, Chelsea J.-T. Ju, Guangyu Zhou, Xuelu Chen, Tianran Zhang, Kai-Wei Chang, Carlo Zaniolo, and Wei Wang. Multifaceted protein-protein interaction prediction based on Siamese residual RCNN. *Bioinformatics (Oxford, England)*, 35(14):i305–i314, July 2019.

[25] Francois Chollet and others. Keras, 2015. Retrieved from https://github.com/fchollet/keras.

[26] Taco S Cohen. *Equivariant convolutional networks*. PhD thesis, University of Amsterdam, 2021.

[27] David K. Cole, Anna M. Bulek, Garry Dolton, Andrea J. Schauenberg, Barbara Szomolay, William Rittase, Andrew Trimby, Prithiviraj Jothikumar, Anna Fuller, Ania Skowera, Jamie Rossjohn, Cheng Zhu, John J. Miles, Mark Peakman, Linda Wooldridge, Pierre J. Rizkallah, and Andrew K. Sewell. Hotspot autoimmune T cell receptor binding underlies pathogen and insulin peptide cross-reactivity. *The Journal of Clinical Investigation*, 126(6):2191–2204, June 2016.

[28] Frances Crawford, Eric Huseby, Janice White, Philippa Marrack, and John W. Kappler. Mimotopes for Alloreactive and Conventional T Cells in a Peptide–MHC Display Library. *PLOS Biology*, 2(4):e90, April 2004. Publisher: Public Library of Science.

[29] G. Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems*, 2(4):303–314, December 1989.

[30] S. Dao-Pin, D. E. Anderson, W. A. Baase, F. W. Dahlquist, and Brian W. Matthews. Structural and thermodynamic consequences of burying a charged residue within the hydrophobic core of T4 lysozyme. *Biochemistry*, 30(49):11521–11529, December 1991.

[31] Mark M. Davis, J. Jay Boniface, Ziv Reich, Daniel Lyons, Johannes Hampl, Bernhard Arden, and Yueh-hsiu Chien. LIGAND RECOGNITION BY $\alpha\beta$ T CELL RECEPTORS. *Annual Review of Immunology*, 16(1):523–544, April 1998.

[32] Juan de Vicente, Juan Lanchares, and Román Hermida. Placement by thermodynamic simulated annealing. *Physics Letters A*, 317(5):415–423, October 2003.

[33] M. M. Dixon, H. Nicholson, L. Shewchuk, W. A. Baase, and B. W. Matthews. Structure of a hinge-bending bacteriophage T4 lysozyme mutant, Ile3 → Pro. *Journal of Molecular Biology*, 227(3):917–933, October 1992.

[34] Yilun Du, Joshua Meier, Jerry Ma, Rob Fergus, and Alexander Rives. Energy-based models for atomic-resolution protein conformations. In *Eighth International Conference on Learning Representations*, April 2020.

[35] Steven M. Dunn, Pierre J. Rizkallah, Emma Baston, Tara Mahon, Brian Cameron, Ruth Moysey, Feng Gao, Malkit Sami, Jonathan Boulter, Yi Li, and Bent K. Jakobsen. Directed evolution of human T cell receptor CDR2 residues by phage display dramatically enhances affinity for cognate peptide-MHC without increasing apparent cross-reactivity. *Protein Science*, 15(4):710–721, 2006. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1110/ps.051936406.

[36] H. Jane Dyson, Peter E. Wright, and Harold A. Scheraga. The role of hydrophobic interactions in initiation and propagation of protein folding. *Proceedings of the National Academy of Sciences*, 103(35):13057–13061, August 2006. Publisher: Proceedings of the National Academy of Sciences.

[37] Ali Esfandiary and Soudeh Ghafouri-Fard. New York esophageal squamous cell carcinoma-1 and cancer immunotherapy. *Immunotherapy*, 7(4):411–439, 2015.

[38] Brian D. Evavold, Joanne Sloan-Lancastert, K. Jeff Wilson, Jonathan B. Rothbard, and Paul M. Allen. Specific T cell recognition of minimally homologous peptides: Evidence for multiple endogenous Ligands. *Immunity*, 2(6):655–663, June 1995.

[39] Alan R. Fersht, Jian-Ping Shi, Jack Knill-Jones, Denise M. Lowe, Anthony J. Wilkinson, David M. Blow, Peter Brick, Paul Carter, Mary M. Y. Waye, and Greg Winter. Hydrogen bonding and biological specificity analysed by protein engineering. *Nature*, 314(6008):235–238, March 1985. Number: 6008 Publisher: Nature Publishing Group.

[40] David S Fischer, Yihan Wu, Benjamin Schubert, and Fabian J Theis. Predicting antigen specificity of single T cells based on TCR CDR3 regions. *Molecular Systems Biology*, 16(8):e9416, August 2020. Publisher: John Wiley & Sons, Ltd.

[41] Travis Gallagher, Patrick Alexander, Philip Bryan, and Gary L. Gilliland. Two Crystal Structures of the B1 Immunoglobulin-Binding Domain of Streptococcal Protein G and Comparison with NMR. *Biochemistry*, 33(15):4721–4729, April 1994. Publisher: American Chemical Society.

[42] D. N. Garboczi, P. Ghosh, U. Utz, Q. R. Fan, W. E. Biddison, and D. C. Wiley. Structure of the complex between human T-cell receptor, viral peptide and HLA-A2. *Nature*, 384(6605):134–141, November 1996.

[43] Nadine C. Gassner, Walter A. Baase, Joel D. Lindstrom, Jirong Lu, Frederick W. Dahlquist, and Brian W. Matthews. Methionine and Alanine Substitutions Show That the Formation of Wild-Type-like Structure in the Carboxy-Terminal Domain of T4 Lysozyme Is a Rate-Limiting Step in Folding. *Biochemistry*, 38(44):14451–14460, November 1999. Publisher: American Chemical Society.

[44] Marvin H. Gee, Arnold Han, Shane M. Lofgren, John F. Beausang, Juan L. Mendoza, Michael E. Birnbaum, Michael T. Bethune, Suzanne Fischer, Xinbo Yang, Raquel Gomez-Eerland, David B. Bingham, Leah V. Sibener, Ricardo A. Fernandes, Andrew Velasco, David Baltimore, Ton N. Schumacher, Purvesh Khatri, Stephen R. Quake, Mark M. Davis, and K. Christopher Garcia. Antigen identification for orphan T cell receptors expressed on tumor-infiltrating lymphocytes. *Cell*, 172(3):549–563.e16, January 2018.

[45] Sofie Gielis, Pieter Moris, Wout Bittremieux, Nicolas De Neuter, Benson Ogunjimi, Kris Laukens, and Pieter Meysman. Detection of Enriched T Cell Epitope Specificity in Full T Cell Receptor Sequence Repertoires. *Frontiers in Immunology*, 10, 2019.

[46] Jacob Glanville, Huang Huang, Allison Nau, Olivia Hatton, Lisa E. Wagar, Florian Rubelt, Xuhuai Ji, Arnold Han, Sheri M. Krams, Christina Pettus, Nikhil Haas, Cecilia S. Lindestam Arlehamn, Alessandro Sette, Scott D. Boyd, Thomas J. Scriba, Olivia M. Martinez, and Mark M. Davis. Identifying specificity groups in the T cell receptor repertoire. *Nature*, 547(7661):94–98, July 2017. Number: 7661 Publisher: Nature Publishing Group.

[47] Ragul Gowthaman and Brian G. Pierce. TCR3d: The T cell receptor structural repertoire database. *Bioinformatics (Oxford, England)*, 35(24):5323–5325, December 2019.

[48] Will Grathwohl, Kuan-Chieh Wang, Joern-Henrik Jacobsen, David Duvenaud, Mohammad Norouzi, and Kevin Swersky. Your classifier is secretly an energy based model and you should treat it like one. In *International Conference on Learning Representations*, 2020.

[49] T M Gray and B W Matthews. Structural analysis of the temperature-sensitive mutant of bacteriophage T4 lysozyme, glycine 156→aspartic acid. *Journal of Biological Chemistry*, 262(35):16858–16864, December 1987.

[50] M. G. Grütter, T. M. Gray, L. H. Weaver, T. Alber, K. Wilson, and B. W. Matthews. Structural studies of mutants of the lysozyme of bacteriophage T4: The temperature-sensitive mutant protein Thr157 → Ile. *Journal of Molecular Biology*, 197(2):315–329, September 1987.

[51] Zhongliang Guo and Rui Yamaguchi. Machine learning methods for protein-protein binding affinity prediction in protein design. *Frontiers in Bioinformatics*, 2, 2022. Publisher: Frontiers Media SA.

[52] W. K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109, April 1970.

[53] David Heckmann, Colton J. Lloyd, Nathan Mih, Yuanchi Ha, Daniel C. Zielinski, Zachary B. Haiman, Abdelmoneim Amer Desouki, Martin J. Lercher, and Bernhard O. Palsson. Machine learning applied to enzyme turnover numbers reveals protein structural correlates and improves metabolic models. *Nature Communications*, 9(1):5252, December 2018.

[54] C. Hélène. Role of aromatic amino-acid residues in the binding of enzymes and proteins to nucleic acids. *Nature: New Biology*, 234(47):120–121, November 1971.

[55] Monica Hollstein, David Sidransky, Bert Vogelstein, and Curtis C. Harris. p53 Mutations in Human Cancers. *Science*, 253(5015):49–53, July 1991. Publisher: American Association for the Advancement of Science.

[56] Thomas A Hopf, Anna G Green, Benjamin Schubert, Sophia Mersmann, Charlotta P I Schärfe, John B Ingraham, Agnes Toth-Petroczy, Kelly Brock, Adam J Riesselman, Perry Palmedo, Chan Kang, Robert Sheridan, Eli J Draizen, Christian Dallago, Chris Sander, and Debora S Marks. The EVcouplings Python framework for coevolutionary sequence analysis. *Bioinformatics*, 35(9):1582–1584, May 2019.

[57] Parisa Hosseinzadeh, Gaurav Bhardwaj, Vikram Khipple Mulligan, Matthew D. Shortridge, Timothy W. Craven, Fátima Pardo-Avila, Stephen A. Rettie, David E. Kim, Daniel-Adriano Silva, Yehia M. Ibrahim, Ian K. Webb, John R. Cort, Joshua N. Adkins, Gabriele Varani, and David Baker. Comprehensive computational design of ordered peptide macrocycles. *Science*, 358(6369):1461–1466, December 2017. Publisher: American Association for the Advancement of Science.

[58] J. A. Hunt and V. M. Ingram. Allelomorphism and the Chemical Differences of the Human Hæmoglobins A, S and C. *Nature*, 181(4615):1062–1063, April 1958. Number: 4615 Publisher: Nature Publishing Group.

[59] James H. Hurley, Walter A. Baase, and Brian W. Matthews. Design and structural analysis of alternative hydrophobic core packing arrangements in bacteriophage T4 lysozyme. *Journal of Molecular Biology*, 224(4):1143–1159, April 1992.

[60] Aapo Hyvärinen. Estimation of Non-Normalized Statistical Models by Score Matching. *Journal of Machine Learning Research*, 6(24):695–709, 2005.

[61] V. M. Ingram. Gene Mutations in Human Hæmoglobin: the Chemical Difference Between Normal and Sickle Cell Hæmoglobin. *Nature*, 180(4581):326–328, August 1957. Number: 4581 Publisher: Nature Publishing Group.

[62] Jeffrey Ishizuka, Kristie Grebe, Eugene Shenderov, Bjoern Peters, Qiongyu Chen, YanChun Peng, Lili Wang, Tao Dong, Valerie Pasquetto, Carla Oseroff, John Sidney, Heather Hickman, Vincenzo Cerundolo, Alessandro Sette, Jack R. Bennink, Andrew McMichael, and Jonathan W. Yewdell. Quantitating T Cell Cross-Reactivity for Unrelated Peptide Antigens1. *The Journal of Immunology*, 183(7):4337–4345, October 2009.

[63] Laura A. Johnson, Richard A. Morgan, Mark E. Dudley, Lydie Cassard, James C. Yang, Marybeth S. Hughes, Udai S. Kammula, Richard E. Royal, Richard M. Sherry, John R. Wunderlich, Chyi-Chia R. Lee, Nicholas P. Restifo, Susan L. Schwarz, Alexandria P. Cogdill, Rachel J. Bishop, Hung Kim, Carmen C. Brewer, Susan F. Rudy, Carter VanWaes, Jeremy L. Davis, Aarti Mathur, Robert T. Ripley, Debbie A. Nathan, Carolyn M. Laurencot, and Steven A. Rosenberg. Gene therapy with human and mouse T-cell receptors mediates cancer regression and targets normal tissues expressing cognate antigen. *Blood*, 114(3):535–546, July 2009.

[64] Emmi Jokinen, Jani Huuhtanen, Satu Mustjoki, Markus Heinonen, and Harri Lähdesmäki. Predicting recognition between T cell receptors and epitopes with TCRGP. *PLOS Computational Biology*, 17(3):e1008814, March 2021. Publisher: Public Library of Science.

[65] Robbie P. Joosten, Jean Salzemann, Vincent Bloch, Heinz Stockinger, Ann-Charlott Berglund, Christophe Blanchet, Erik Bongcam-Rudloff, Christophe Combet, Ana L. Da Costa, Gilbert Deleage, Matteo Diarena, Roberto Fabbretti, Géraldine Fettahi, Volker Flegel, Andreas Gisel, Vinod Kasam, Timo Kervinen, Eija Korpelainen, Kimmo Mattila, Marco Pagni, Matthieu Reichstadt, Vincent Breton, Ian J. Tickle, and Gert Vriend. PDB_redo: automated re-refinement of X-ray structure models in the PDB. *Journal of Applied Crystallography*, 42(Pt 3):376–384, April 2009.

[66] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A. A. Kohl, Andrew J. Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W. Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589, August 2021.

[67] Thomas Kammertoens and Thomas Blankenstein. It's the Peptide-MHC Affinity, Stupid. *Cancer Cell*, 23(4):429–431, April 2013.

[68] Panagiotis L. Kastritis and Alexandre M. J. J. Bonvin. On the binding affinity of macromolecular interactions: daring to ask why proteins interact. *Journal of the Royal Society Interface*, 10(79):20120835, February 2013.

[69] Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[70] Risi Kondor, Zhen Lin, and Shubhendu Trivedi. Clebsch–Gordan nets: a fully fourier space spherical convolutional neural network. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, pages 10138–10147, Red Hook, NY, USA, December 2018. Curran Associates Inc.

[71] Jan Kubelka, James Hofrichter, and William A Eaton. The protein folding 'speed limit'. *Current Opinion in Structural Biology*, 14(1):76–88, February 2004.

[72] Nicole L. La Gruta, Stephanie Gras, Stephen R. Daley, Paul G. Thomas, and Jamie Rossjohn. Understanding the drivers of MHC restriction of T cell receptors. *Nature Reviews Immunology*, 18(7):467–478, July 2018. Number: 7 Publisher: Nature Publishing Group.

[73] Jun Lan, Jiwan Ge, Jinfang Yu, Sisi Shan, Huan Zhou, Shilong Fan, Qi Zhang, Xuanling Shi, Qisheng Wang, Linqi Zhang, and Xinquan Wang. Structure of the SARS-CoV-2 spike receptor-binding domain bound to the ACE2 receptor. *Nature*, 581(7807):215–220, May 2020. Number: 7807 Publisher: Nature Publishing Group.

[74] Moshe Leshno, Vladimir Ya. Lin, Allan Pinkus, and Shimon Schocken. Multilayer feedforward networks with a nonpolynomial activation function can approximate any function. *Neural Networks*, 6(6):861–867, January 1993.

[75] Feiran Li, Le Yuan, Hongzhong Lu, Gang Li, Yu Chen, Martin K. M. Engqvist, Eduard J. Kerkhoven, and Jens Nielsen. Deep learning-based kcat prediction enables improved enzyme-constrained model reconstruction. *Nature Catalysis*, 5(8):662–672, August 2022. Number: 8 Publisher: Nature Publishing Group.

[76] Yi Li, Ruth Moysey, Peter E. Molloy, Anne-Lise Vuidepot, Tara Mahon, Emma Baston, Steven Dunn, Nathaniel Liddy, Jansen Jacob, Bent K. Jakobsen, and Jonathan M. Boulter. Directed evolution of human T-cell receptors with picomolar affinities by phage display. *Nature Biotechnology*, 23(3):349–354, March 2005. Number: 3 Publisher: Nature Publishing Group.

[77] Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Salvatore Candido, and Alexander Rives. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, March 2023. Publisher: American Association for the Advancement of Science.

[78] Sara Linse, Peter Brodin, Charlotta Johansson, Eva Thulin, Thomas Grundström, and Sture Forsén. The role of protein surface charges in ion binding. *Nature*, 335(6191):651–652, October 1988. Number: 6191 Publisher: Nature Publishing Group.

[79] Leigh Ann Lipscomb, Nadine C. Gassner, Sheila D. Snow, Aimee M. Eldridge, Walter A. Baase, Devin L. Drew, and Brain W. Matthews. Context-dependent protein stabilization by methionine-to-leucine substitution shown in T4 lysozyme. *Protein Science*, 7(3):765–773, 1998. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/pro.5560070326.

[80] Sidney Lyayuga Lisanza, Jake Merle Gershon, Sam Tipps, Lucas Arnoldt, Samuel Hendel, Jeremiah Nelson Sims, Xinting Li, and David Baker. Joint Generation of Protein Sequence and Structure with RoseTTAFold Sequence Space Diffusion, May 2023. Pages: 2023.05.08.539766 Section: New Results.

[81] Xianggen Liu, Yunan Luo, Pengyong Li, Sen Song, and Jian Peng. Deep geometric representations for modeling effects of mutations on protein-protein binding affinity. *PLoS computational biology*, 17(8):e1009284, August 2021.

[82] Justin M. Long and David M. Holtzman. Alzheimer Disease: An Update on Pathobiology and Treatment Strategies. *Cell*, 179(2):312–339, October 2019.

[83] Simon C. Lovell, Ian W. Davis, W. Bryan Arendall III, Paul I. W. de Bakker, J. Michael Word, Michael G. Prisant, Jane S. Richardson, and David C. Richardson. Structure validation by Cα geometry: ϕ,ψ and Cβ deviation. *Proteins: Structure, Function, and Bioinformatics*, 50(3):437–450, 2003. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/prot.10286.

[84] Masazumi Matsumura, Wayne J. Becktel, and Brian W. Matthews. Hydrophobic stabilization in T4 lysozyme determined directly by multiple substitutions of Ile 3. *Nature*, 334(6181):406–410, August 1988. Number: 6181 Publisher: Nature Publishing Group.

[85] B W Matthews, H Nicholson, and W J Becktel. Enhanced protein thermostability from site-directed mutations that decrease the entropy of unfolding. *Proceedings of the National Academy of Sciences of the United States of America*, 84(19):6663–6667, October 1987.

[86] Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger. UMAP: Uniform Manifold Approximation and Projection. *Journal of Open Source Software*, 3(29):861, September 2018.

[87] Nicholas Metropolis, Arianna W. Rosenbluth, Marshall N. Rosenbluth, Augusta H. Teller, and Edward Teller. Equation of State Calculations by Fast Computing Machines. *The Journal of Chemical Physics*, 21(6):1087–1092, June 1953.

[88] Martina Milighetti, John Shawe-Taylor, and Benny Chain. Predicting T Cell Receptor Antigen Specificity From Structural Features Derived From Homology Models of Receptor-Peptide-Major Histocompatibility Complexes. *Frontiers in Physiology*, 12, 2021.

[89] Blaine H. M. Mooers, Walter A. Baase, Jonathan W. Wray, and Brian W. Matthews. Contributions of all 20 amino acids at site 96 to the stability and structure of T4 lysozyme. *Protein Science*, 18(5):871–880, May 2009.

[90] Blaine H. M. Mooers, Deepshikha Datta, Walter A. Baase, Eric S. Zollars, Stephen L. Mayo, and Brian W. Matthews. Repacking the Core of T4 Lysozyme by Automated Design. *Journal of Molecular Biology*, 332(3):741–756, September 2003.

[91] Faruck Morcos, Andrea Pagnani, Bryan Lunt, Arianna Bertolino, Debora S. Marks, Chris Sander, Riccardo Zecchina, José N. Onuchic, Terence Hwa, and Martin Weigt. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proceedings of the National Academy of Sciences*, 108(49):E1293–E1301, December 2011. Publisher: Proceedings of the National Academy of Sciences.

[92] Pieter Moris, Joey De Pauw, Anna Postovskaya, Sofie Gielis, Nicolas De Neuter, Wout Bittremieux, Benson Ogunjimi, Kris Laukens, and Pieter Meysman. Current challenges for unseen-epitope TCR interaction prediction and a new perspective derived from image classification. *Briefings in Bioinformatics*, 22(4):bbaa318, July 2021.

[93] Albert Musaelian, Simon Batzner, Anders Johansson, Lixin Sun, Cameron J. Owen, Mordechai Kornbluth, and Boris Kozinsky. Learning local equivariant representations for large-scale atomistic dynamics. *Nature Communications*, 14:579, February 2023.

[94] H. Nicholson, D. E. Anderson, S. Dao-pin, and B. W. Matthews. Analysis of the interaction between charged side chains and the alpha-helix dipole using designed thermostable mutants of phage T4 lysozyme. *Biochemistry*, 30(41):9816–9828, October 1991.

[95] H. Nicholson, W. J. Becktel, and B. W. Matthews. Enhanced protein thermostability from designed mutations that interact with α-helix dipoles. *Nature*, 336(6200):651–656, December 1988. Number: 6200 Publisher: Nature Publishing Group.

[96] Zaneta Nikolovska-Coleska. Studying Protein-Protein Interactions Using Surface Plasmon Resonance. In Cheryl L. Meyerkord and Haian Fu, editors, *Protein-Protein Interactions: Methods and Applications*, Methods in Molecular Biology, pages 109–138. Springer, New York, NY, 2015.

[97] C. N. Pace, B. A. Shirley, M. McNutt, and K. Gajiwala. Forces contributing to the conformational stability of proteins. *FASEB journal: official publication of the Federation of American Societies for Experimental Biology*, 10(1):75–83, January 1996.

[98] Sujatha Thankeswaran Parvathy, Varatharajalu Udayasuriyan, and Vijaipal Bhadana. Codon usage bias. *Molecular Biology Reports*, 49(1):539–565, 2022.

[99] Johannes Pettmann, Anna Huhn, Enas Abu Shah, Mikhail A Kutuzov, Daniel B Wilson, Michael L Dustin, Simon J Davis, P Anton van der Merwe, and Omer Dushek. The discriminatory power of the T cell receptor. *eLife*, 10:e67092, May 2021. Publisher: eLife Sciences Publications, Ltd.

[100] Kurt H. Piepenbrink, Brian E. Gloor, Kathryn M. Armstrong, and Brian M. Baker. Methods for quantifying T cell receptor binding affinities and thermodynamics. *Methods in Enzymology*, 466:359–381, 2009.

[101] Novalia Pishesha, Thibault J. Harmand, and Hidde L. Ploegh. A guide to antigen processing and presentation. *Nature Reviews Immunology*, 22(12):751–764, December 2022. Number: 12 Publisher: Nature Publishing Group.

[102] Philip Pjura and Brian W. Matthews. Structures of randomly generated mutants of T4 lysozyme show that protein stability can be enhanced by relaxation of strain and by improved hydrogen bonding via bound solvent. *Protein Science*, 2(12):2226–2232, 1993. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/pro.5560021222.

[103] Christine Tran Quang, Benedetta Zaniboni, and Jacques Ghysdael. A TCR-switchable cell death pathway in T-ALL. *Oncoscience*, 4(3-4):17–18, March 2017.

[104] Timothy P. Riley, Lance M. Hellman, Marvin H. Gee, Juan L. Mendoza, Jesus A. Alonso, Kendra C. Foley, Michael I. Nishimura, Craig W. Vander Kooi, K. Christopher Garcia, and Brian M. Baker. T cell receptor cross-reactivity expanded by dramatic peptide–MHC adaptability. *Nature Chemical Biology*, 14(10):934–942, October 2018. Number: 10 Publisher: Nature Publishing Group.

[105] Serina L Robinson, Megan D Smith, Jack E Richman, Kelly G Aukema, and Lawrence P Wackett. Machine learning-based prediction of activity and substrate specificity for OleA enzymes in the thiolase superfamily. *Synthetic Biology*, 5(1):ysaa004, January 2020.

[106] Carlos H. M. Rodrigues, Douglas E. V. Pires, and David B. Ascher. mmCSM-PPI: predicting the effects of multiple point mutations on protein-protein interactions. *Nucleic Acids Research*, 49(W1):W417–W424, July 2021.

[107] G. C. Román and M. Osame. Identity of HTLV-I-associated tropical spastic paraparesis and HTLV-I-associated myelopathy. *Lancet (London, England)*, 1(8586):651, March 1988.

[108] Pedro Romero, Danila Valmori, Mikael J. Pittet, Alfred Zippelius, Donata Rimoldi, Frederic Lévy, Valérie Dutoit, Maha Ayyoub, Verena Rubio-Godoy, Olivier Michielin, Philippe Guillaume, Pascal Batard, Immanuel F. Luescher, Ferdy Lejeune, Danielle Liénard, Nathalie Rufer, Pierre-Yves Dietrich, Daniel E. Speiser, and Jean-Charles Cerottini. Antigenicity and immunogenicity of Melan-A/MART-1 derived peptides as targets for tumor reactive CTL in human melanoma. *Immunological Reviews*, 188:81–96, October 2002.

[109] Sandra Romero-Molina, Yasser B. Ruiz-Blanco, Joel Mieres-Perez, Mirja Harms, Jan Münch, Michael Ehrmann, and Elsa Sanchez-Garcia. PPI-Affinity: A Web Tool for the Prediction and Optimization of Protein-Peptide and Protein-Protein Binding Affinity. *Journal of Proteome Research*, 21(8):1829–1841, August 2022.

[110] James P. Roney and Sergey Ovchinnikov. State-of-the-Art Estimation of Protein Model Accuracy Using AlphaFold. *Physical Review Letters*, 129(23):238101, November 2022.

[111] Steven A. Rosenberg, Nicholas P. Restifo, James C. Yang, Richard A. Morgan, and Mark E. Dudley. Adoptive cell transfer: a clinical path to effective cancer immunotherapy. *Nature reviews. Cancer*, 8(4):299–308, April 2008.

[112] Jamie Rossjohn, Stephanie Gras, John J. Miles, Stephen J. Turner, Dale I. Godfrey, and James McCluskey. T cell antigen receptor recognition of antigen-presenting molecules. *Annual Review of Immunology*, 33:169–200, 2015.

[113] Michel Sadelain, Renier Brentjens, and Isabelle Rivière. The Basic Principles of Chimeric Antigen Receptor Design. *Cancer Discovery*, 3(4):388–398, April 2013.

[114] Tiziana Sanavia, Giovanni Birolo, Ludovica Montanucci, Paola Turina, Emidio Capriotti, and Piero Fariselli. Limitations and challenges in protein stability prediction upon genome variations: towards future applications in precision medicine. *Computational and Structural Biotechnology Journal*, 18:1968–1979, July 2020.

[115] Schrödinger, LLC. The PyMOL Molecular Graphics System, Version 1.8. November 2015.

[116] Kinjal Shah, Amr Al-Haidari, Jianmin Sun, and Julhash U. Kazi. T cell receptor (TCR) signaling in health and disease. *Signal Transduction and Targeted Therapy*, 6(1):1–26, December 2021. Number: 1 Publisher: Nature Publishing Group.

[117] Raghav Shroff, Austin W. Cole, Daniel J. Diaz, Barrett R. Morrow, Isaac Donnell, Ankur Annapareddy, Jimmy Gollihar, Andrew D. Ellington, and Ross Thyer. Discovery of Novel Gain-of-Function Mutations Guided by Structure-Based Deep Learning. *ACS Synthetic Biology*, 9(11):2927–2935, November 2020. Publisher: American Chemical Society.

[118] John-William Sidhom, H. Benjamin Larman, Drew M. Pardoll, and Alexander S. Baras. DeepTCR is a deep learning framework for revealing sequence concepts within T-cell repertoires. *Nature Communications*, 12(1):1605, March 2021. Number: 1 Publisher: Nature Publishing Group.

[119] Timothy R. Siegert, Michael J. Bird, Kamlesh M. Makwana, and Joshua A. Kritzer. Analysis of Loops that Mediate Protein–Protein Interactions and Translation into Submicromolar Inhibitors. *Journal of the American Chemical Society*, 138(39):12876–12884, October 2016. Publisher: American Chemical Society.

[120] Umut Simsekli. Posterior Sampling with Stochastic Gradient Langevin Dynamics.

[121] Nishant K. Singh, Timothy P. Riley, Sarah Catherine B. Baker, Tyler Borrman, Zhiping Weng, and Brian M. Baker. Emerging Concepts in TCR Specificity: Rationalizing and (Maybe) Predicting Outcomes. *Journal of Immunology (Baltimore, Md.: 1950)*, 199(7):2203–2213, October 2017.

[122] Yang Song and Stefano Ermon. Generative Modeling by Estimating Gradients of the Data Distribution. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[123] Carolina M. Soto, Jennifer D. Stone, Adam S. Chervin, Boris Engels, Hans Schreiber, Edward J. Roy, and David M. Kranz. MHC-class I-restricted CD4 T cells: a nanomolar affinity TCR has improved anti-tumor efficacy in vivo compared to the micromolar wild-type TCR. *Cancer Immunology, Immunotherapy*, 62(2):359–369, February 2013.

[124] Tyler N. Starr, Allison J. Greaney, William W. Hannon, Andrea N. Loes, Kevin Hauser, Josh R. Dillen, Elena Ferri, Ariana Ghez Farrell, Bernadeta Dadonaite, Matthew McCallum, Kenneth A. Matreyek, Davide Corti, David Veesler, Gyorgy Snell, and Jesse D. Bloom. Shifting mutational constraints in the SARS-CoV-2 receptor-binding domain during viral evolution. *Science*, 377(6604):420–424, July 2022. Publisher: American Association for the Advancement of Science.

[125] Tyler N. Starr, Allison J. Greaney, Sarah K. Hilton, Daniel Ellis, Katharine H. D. Crawford, Adam S. Dingens, Mary Jane Navarro, John E. Bowen, M. Alejandra Tortorici, Alexandra C. Walls, Neil P. King, David Veesler, and Jesse D. Bloom. Deep Mutational Scanning of SARS-CoV-2 Receptor Binding Domain Reveals Constraints on Folding and ACE2 Binding. *Cell*, 182(5):1295–1310.e20, September 2020.

[126] Matteo Stefanini, Marta Lovino, Rita Cucchiara, and Elisa Ficarra. Predicting gene and protein expression levels from DNA and protein sequences with Perceiver. *Computer Methods and Programs in Biomedicine*, 234:107504, June 2023.

[127] Robert C. Sterner and Rosalie M. Sterner. CAR-T cell therapy: current limitations and potential strategies. *Blood Cancer Journal*, 11(4):69, April 2021.

[128] Elliott J Stollar and David P Smith. Uncovering protein structure. *Essays in Biochemistry*, 64(4):649–680, September 2020.

[129] Jennifer D Stone, Adam S Chervin, and David M Kranz. T-cell receptor binding affinities and kinetics: impact on T-cell activity and specificity. *Immunology*, 126(2):165–176, February 2009.

[130] Jennifer D. Stone and David M. Kranz. Role of T Cell Receptor Affinity in the Efficacy and Specificity of Adoptive T Cell Therapies. *Frontiers in Immunology*, 4:244, August 2013.

[131] Jan Stourac, Juraj Dubrava, Milos Musil, Jana Horackova, Jiri Damborsky, Stanislav Mazurenko, and David Bednar. FireProtDB: database of manually curated protein stability data. *Nucleic Acids Research*, 49(D1):D319–D324, January 2021.

[132] Joonho Suh and Yun-Sil Lee. Similar sequences but dissimilar biological functions of GDF11 and myostatin. *Experimental & Molecular Medicine*, 52(10):1673–1693, October 2020. Number: 10 Publisher: Nature Publishing Group.

[133] Sylvanie Surget, Marie P. Khoury, and Jean-Christophe Bourdon. Uncovering the role of p53 splice variants in human malignancy: a clinical perspective. *OncoTargets and therapy*, 7:57, 2014. Publisher: Dove Press.

[134] Baris E. Suzek, Hongzhan Huang, Peter McGarvey, Raja Mazumder, and Cathy H. Wu. UniRef: comprehensive and non-redundant UniProt reference clusters. *Bioinformatics*, 23(10):1282–1288, May 2007.

[135] Nathaniel Thomas, Tess Smidt, Steven Kearnes, Lusann Yang, Li Li, Kai Kohlhoff, and Patrick Riley. Tensor field networks: Rotation- and translation-equivariant neural networks for 3D point clouds, May 2018. arXiv:1802.08219 [cs].

[136] Paul G. Thomas, Alan J. Russell, and Alan R. Fersht. Tailoring the pH dependence of enzyme catalysis using protein engineering. *Nature*, 318(6044):375–376, November 1985. Number: 6044 Publisher: Nature Publishing Group.

[137] Wen Torng and Russ B. Altman. 3D deep convolutional neural networks for amino acid environment similarity analysis. *BMC Bioinformatics*, 18(1):302, June 2017.

[138] Wu-Ki Tung. *Group theory in physics*. World Scientific, Philadelphia, 1985.

[139] U. Utz, D. Banks, S. Jacobson, and W. E. Biddison. Analysis of the T-cell receptor repertoire of human T-cell leukemia virus type 1 (HTLV-1) Tax-specific CD8+ cytotoxic T lymphocytes from patients with HTLV-1-associated disease: evidence for oligoclonal expansion. *Journal of Virology*, 70(2):843–851, February 1996.

[140] Anna Vangone and Alexandre Mjj Bonvin. Contacts-based prediction of binding affinity in protein-protein complexes. *eLife*, 4:e07454, July 2015.

[141] Ryan G. Walker, Tommaso Poggioli, Lida Katsimpardi, Sean M. Buchanan, Juhyun Oh, Sam Wattrus, Bettina Heidecker, Yick W. Fong, Lee L. Rubin, Peter Ganz, Thomas B. Thompson, Amy J. Wagers, and Richard T. Lee. Biochemistry and Biology of GDF11 and Myostatin. *Circulation Research*, 118(7):1125–1142, April 2016. Publisher: American Heart Association.

[142] Menglun Wang, Zixuan Cang, and Guo-Wei Wei. A topology-based network tree for the prediction of protein-protein binding affinity changes following mutation. *Nature Machine Intelligence*, 2(2):116–123, 2020.

[143] Yiquan Wang, Ruipeng Lei, Armita Nourmohammad, and Nicholas C. Wu. Antigenic evolution of human influenza H3N2 neuraminidase is constrained by charge balancing. *eLife*, 10:e72516, December 2021.

[144] Stephen G. Ward and Carl H. June. T Lymphocyte Activation. In Peter J. Delves, editor, *Encyclopedia of Immunology (Second Edition)*, pages 2323–2329. Elsevier, Oxford, January 1998.

[145] L. H. Weaver, T. M. Gray, M. G. Grütter, D. E. Anderson, J. A. Wozniak, F. W. Dahlquist, and B. W. Matthews. High-resolution structure of the temperature-sensitive mutant of phage lysozyme, Arg 96—-His. *Biochemistry*, 28(9):3793–3797, May 1989.

[146] L. H. Weaver and B. W. Matthews. Structure of bacteriophage T4 lysozyme refined at 1.7 å resolution. *Journal of Molecular Biology*, 193(1):189–199, January 1987.

[147] Anna Weber, Jannis Born, and María Rodriguez Martínez. TITAN: T-cell receptor specificity prediction with bimodal attention networks. *Bioinformatics*, 37(Supplement_1):i237–i244, July 2021.

[148] JunJie Wee and Kelin Xia. Persistent spectral based ensemble learning (PerSpect-EL) for protein-protein binding affinity prediction. *Briefings in Bioinformatics*, 23(2):bbac024, March 2022.

[149] Martin Weigt, Robert A. White, Hendrik Szurmant, James A. Hoch, and Terence Hwa. Identification of direct residue contacts in protein–protein interaction by message passing. *Proceedings of the National Academy of Sciences*, 106(1):67–72, January 2009. Publisher: Proceedings of the National Academy of Sciences.

[150] Maurice Weiler, Mario Geiger, Max Welling, Wouter Boomsma, and Taco Cohen. 3D steerable CNNs: learning rotationally equivariant features in volumetric data. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, pages 10402–10413, Red Hook, NY, USA, December 2018. Curran Associates Inc.

[151] Max Welling and Yee Whye Teh. Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, ICML'11, pages 681–688, Madison, WI, USA, June 2011. Omnipress.

[152] David Whitford. *Proteins: structure and function*. J. Wiley & Sons, Hoboken, NJ, 2005.

[153] Darcy B Wilson, Dianne H Wilson, Kim Schroder, Clemencia Pinilla, Sylvie Blondelle, Richard A Houghten, and K. Christopher Garcia. Specificity and degeneracy of T cells. *Molecular Immunology*, 40(14):1047–1055, February 2004.

[154] Eric T. C. Wong, Dokyun Na, and Jörg Gsponer. On the Importance of Polar Interactions for Complexes Containing Intrinsically Disordered Proteins. *PLOS Computational Biology*, 9(8):e1003192, August 2013. Publisher: Public Library of Science.

[155] Linda Wooldridge, Julia Ekeruche-Makinde, Hugo A. van den Berg, Anna Skowera, John J. Miles, Mai Ping Tan, Garry Dolton, Mathew Clement, Sian Llewellyn-Lacey, David A. Price, Mark Peakman, and Andrew K. Sewell. A single autoimmune T cell receptor recognizes more than a million different peptides. *The Journal of Biological Chemistry*, 287(2):1168–1177, January 2012.

[156] D C Wraith, B Bruun, and P J Fairchild. Cross-reactive antigen recognition by an encephalitogenic T cell receptor. Implications for T cell biology and autoimmunity. *The Journal of Immunology*, 149(11):3765–3770, December 1992.

[157] Jonathan W Wray, Walter A Baase, Joel D Lindstrom, Larry H Weaver, Anthony R Poteete, and Brian W Matthews. Structural analysis of a non-contiguous second-site revertant in T4 lysozyme shows that increasing the rigidity of a protein can enhance its stability11Edited by J. A. Wells. *Journal of Molecular Biology*, 292(5):1111–1120, October 1999.

[158] Nicholas C. Wu, Jakub Otwinowski, Andrew J. Thompson, Corwin M. Nycholat, Armita Nourmohammad, and Ian A. Wilson. Major antigenic site B of human influenza H3N2 viruses has an evolving local fitness landscape. *Nature Communications*, 11(1):1233, March 2020. Number: 1 Publisher: Nature Publishing Group.

[159] Jian Xu, Walter A. Baase, Michael L. Quillin, Enoch P. Baldwin, and Brian W. Matthews. Structural and thermodynamic analysis of the binding of solvent at internal sites in T4 lysozyme. *Protein Science*, 10(5):1067–1078, 2001. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1110/ps.02101.

[160] Danlin Yang, Rachel Kroe-Barrett, Sanjaya Singh, and Thomas Laue. IgG Charge: Practical and Biological Implications. *Antibodies*, 8(1):24, March 2019.

[161] Xinbo Yang, Lee I. Garner, Ivan V. Zvyagin, Michael A. Paley, Ekaterina A. Komech, Kevin M. Jude, Xiang Zhao, Ricardo A. Fernandes, Lynn M. Hassman, Grace L. Paley, Christina S. Savvides, Simon Brackenridge, Max N. Quastel, Dmitriy M. Chudakov, Paul Bowness, Wayne M. Yokoyama, Andrew J. McMichael, Geraldine M. Gillespie, and K. Christopher Garcia. Autoimmunity-associated T cell receptors recognize HLA-B*27-bound peptides. *Nature*, pages 1–7, December 2022. Publisher: Nature Publishing Group.

[162] K. Yugandhar and M. Michael Gromiha. Protein-protein binding affinity prediction from amino acid sequence. *Bioinformatics (Oxford, England)*, 30(24):3583–3589, December 2014.

[163] Shuqi Zhang, Patricia Parker, Keyue Ma, Chenfeng He, Qian Shi, Zhonghao Cui, Chad Williams, Ben S. Wendel, Amanda Meriwether, Mary A. Salazar, and Ning Jiang. Direct Measurement of T Cell Receptor Affinity and Sequence from Naïve Anti-Viral T Cells. *Science translational medicine*, 8(341):341ra77, June 2016.

[164] Lingling Zhao, Yan Zhu, Junjie Wang, Naifeng Wen, Chunyu Wang, and Liang Cheng. A brief review of protein–ligand interaction prediction. *Computational and Structural Biotechnology Journal*, 20:2831–2838, June 2022.

[165] Simiao Zhao. Prediction of Protein Expression and Growth Rates by Supervised Machine Learning. *Natural Science*, 13(8):301–330, August 2021. Number: 8 Publisher: Scientific Research Publishing.

[166] Shi Zhong, Karolina Malecek, Laura A. Johnson, Zhiya Yu, Eleazar Vega-Saenz de Miera, Farbod Darvishian, Katelyn McGary, Kevin Huang, Josh Boyer, Emily Corse, Yongzhao Shao, Steven A. Rosenberg, Nicholas P. Restifo, Iman Osman, and Michelle Krogsgaard. T-cell receptor affinity and avidity defines antitumor response and autoimmunity in T-cell immunotherapy. *Proceedings of the National Academy of Sciences*, 110(17):6973–6978, April 2013. Publisher: Proceedings of the National Academy of Sciences.

Appendix A

SUPPLEMENT FOR CHAPTER 2

A.1 Code and data availability.

All codes and references to data are available on github through: https://github.com/StatPhysBio/protein_holography.

A.2 Training, validation, and test datasets for amino acid neighborhood classification

We use ProteinNet's splittings for training and validation sets of the CASP12 competition, to avoid redundancy due to similarities in homologous protein domains [5]. Since PDB identifiers in ProteinNet were only provided for the training and the validation sets, we used ProteinNet's training set as the base of both our training and validation, and ProteinNet's validation set as our testing set. Specifically, define our training and validation sets, we make a [80%, 20%] split in the ProteinNet's training data of proteins with X-ray crystallography structures of 2.5Å resolution or better. Furthermore, in anticipation of validating our model on experimental stabilities of T4 Lysozyme and SARS-CoV-2 receptor binding domain (RBD), we removed all structures that belonged to the same UniProt family as the domains from each of these proteins. We used all of ProteinNet's validation structures as our testing set.

This splitting resulted in 10,957 training structures, 2,730 validation structures, and 212 testing structures, each of which contained 2,810,503 training neighborhoods, 682,689 validation neighborhoods, and 4,472 testing neighborhoods.

A.3 Training procedure

Networks were trained with all training data being read at run time using the keras `fit` function and a data generator made from a subclass of the `keras.utils.sequence` class [25]. Generally 40 workers were used for reading the data and a `max_queue_size` of 900 was found to be optimal for read/evaluation balance. We evaluated the model at intervals shorter than one epoch to allow for more fine grained updates to hyperparameters schedules and early stopping. Roughly one tenth of the validation data (51,200 neighborhoods) was evaluated after an approximate one tenth of the training data (256,000 neighborhoods) was seen. This was implemented by using `tf.keras.Model.fit` arguments `steps_per_epoch=1000` and `validation_steps=100`. Data was shuffled at the end of each evaluation interval. These evaluation intervals also served as checkpoints for both early stopping and weight saving. The best network was trained for 8.47 epochs in 4.54 hours on one NVIDIA A40 GPU hosted at the Hyak supercomputer at the University of Washington.

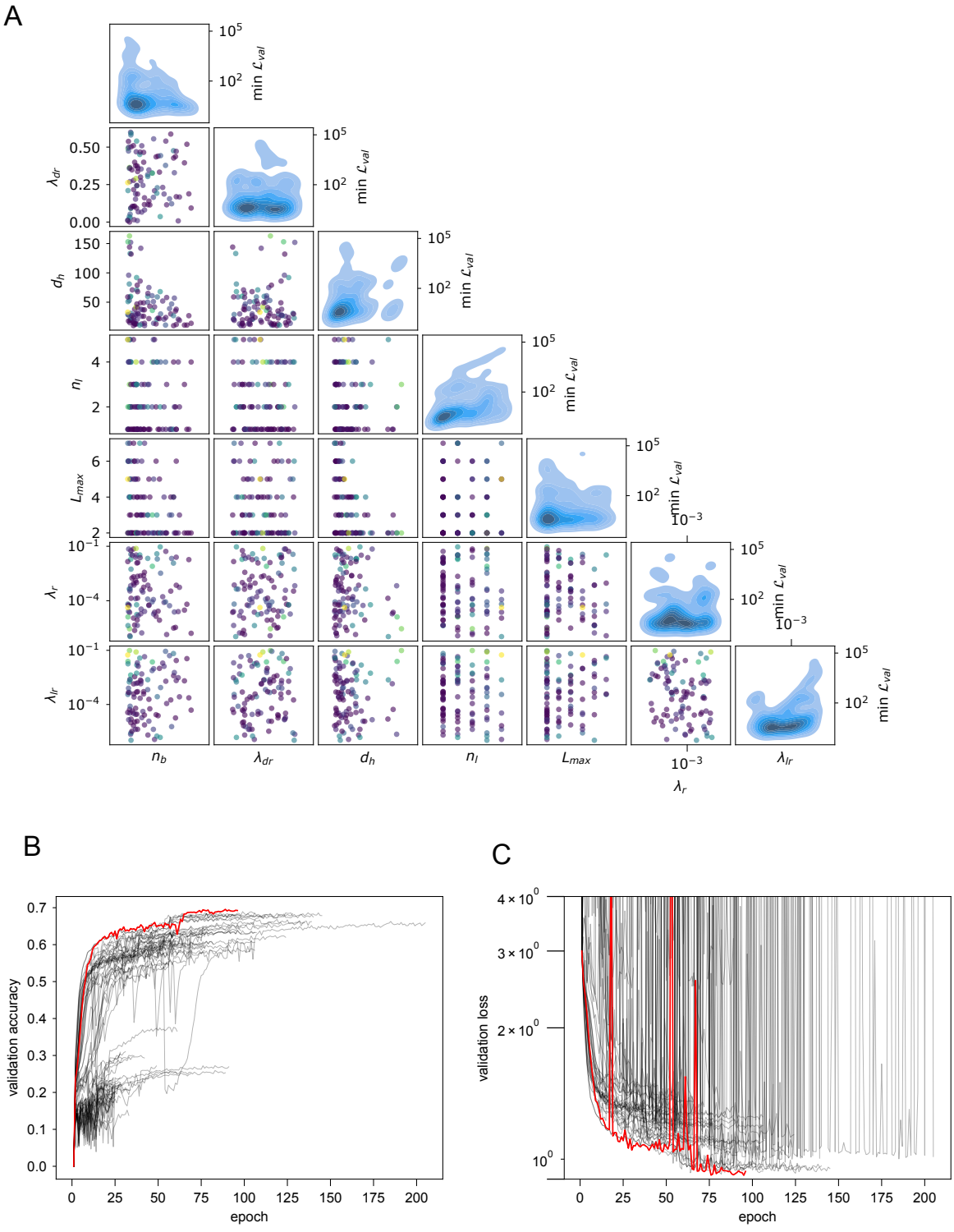
We used stochastic gradient descent and Adam optimizer [69] for training our networks. During development, models trained with Adam generally showed better training and validation loss, and thus, all networks were trained with Adam during hyperparameter optimization. The main component of the loss was defined as the categorical cross-entropy between the one-hot amino acid encoding and the softmax output of the network

$$\mathcal{L} = -\sum_{i=1}^{N} \log P_{\alpha_i} \tag{A.1}$$

where P_{α_i} is the network-assigned probability of the true central amino acid α_i in neighborhood i.

Some hyperparameters were scheduled according to the behavior of the network's loss. The learning rate was put on a schedule of `ReduceLROnPLateau` to decrease by a factor of 10 whenever the validation loss did not improve over 10 tenths of an epoch, with a minimum learning rate of 10^{-9}. Ultimately, the networks were trained with early stopping when the validation loss did not improve by a difference of 0.01 over the course of 20 tenths of an epoch.

Figure A.1: **Hyperparameter optimization for H-CNN.** Hyperparameter optimization was performed in two steps for both the simply connected and the fully connected networks: First, hyperparameters were scanned across a wide regime of hyper-parameters via random sampling. From this scan, a narrow band of hyper-parameters were analyzed to determine the optimal network **(A)**. Off the diagonal, each dot's color shows the validation loss seen during this first step of hyperparameter tuning for a network with the combination of hyperparameters shown on the two axes. On the diagonal, the density of networks is shown as a function of loss and the hyperparameter tested. Validation accuracy **(B)** and loss **(C)** are shown as a function of training epochs for all networks trained in the second stage of hyperparameter tuning. Red lines show the network with the best validation loss that is used throughout this work.


A
B
C
min $\mathcal{L}_{val}$
n_b
λ_{dr}
d_h
n_l
L_{max}
λ_r
λ_{lr}
validation accuracy
validation loss
epoch

A.4 Hyperparameter optimization

Hyperparameter optimization was performed in two steps: First, we scanned a larger parameter regime and trained 500 networks to identify a reasonable subregion of the parameter space (Fig. A.1A). This sampling revealed networks with parameter ranges listed in Table A.1. These parameters were then sampled more finely to get the best network models. The resulted training curves are shown in Fig. A.1B,C.

It should be noted that we optimized over two classes of networks: (i) fully connected networks and (ii) simply connected networks, as defined in Section 2.1.2. For both classes of networks, our hyperparameter optimization scanned the limits of networks that would fit on the GPUs used, however the different choice of connectedness determines the largest component of the network. In simply connected networks, the nonlinear Clebsch-Gordan products are restricted to the square of Fourier coefficients with the same spherical degree ℓ and channel index k. In contrast, fully connected networks use bilinear products between all sets of Fourier coefficients (see equation (2.4)). Thus, simply connected networks can handle more input channels into the nonlinearity while fully connected networks cannot. The trade-off is that fully connected networks produce more output channels via combining all combinations of input channels in the bilinear product (see equations 2.6 & 2.5). Thus the restricted nonlinearities of simply connected networks allowed us to use linear layers that had roughly one order of magnitude larger output dimensions in simply vs. fully connected networks (150 as opposed to 20); see Fig. A.1E. We separately optimized the hyperparameters for these two network classes.

Hyperparameter	variable	broad sampling space	sampling scheme	fine sampling space	optimal fc network value	optimal sc network value
batch size	n_b	[8, 5000]	uniform	256	256	256
dropout rate	λ_{dr}	[0,0.6]	uniform	$[0, 10^{-5}]$	5.49×10^{-4}	1.95×10^{-2}
hidden dimension	d_h	fc:[6,20], sc:[50,160]	uniform	fc:[8,20], sc:[100,500]	14	109
No. CG layers	n_l	[1,5]	uniform	[4,5]	4	5
No. dense layers	n_d	[1,5]	uniform	[2,4]	2	2
Max spherical degree	L_{max}	[2,7]	uniform	[4,5]	5	5
learning rate	λ_l	$[10^{-6}, 10^{-1}]$	exponential	10^{-3}	10^{-3}	10^{-3}
regularization strength	λ_r	$[10^{-6}, 10^{-1}]$	exponential	$[10^{-10}, 10^{-16}]$	1.2×10^{-16}	1.89×10^{-14}

Table A.1: **Hyperparameter bounds and optimal values.** Hyperparameters varied during model optimization are listed. The two different bounds used during the two steps of hyperparameter optimization are shown in the broad and fine sampling spaces. Different bounds for fully connected (fc) and simply connected (sc) networks are shown when appropriate. Finally, the optimal value is listed for both the optimal fully connected and simply connected networks.

Appendix B

SUPPLEMENT FOR CHAPTER 3

B.1 Data for assessing the effect of shearing in Protein G

To analyze the effect of shearing in protein G, shown in Figs. 3.1 and 3.2, we used the native structure with the PDB identifier 1PGA [41]. Sheared structures were produced with PyRosetta by manually editing the backbone angles ϕ and ψ of each residue (Fig. 3.1A) via `Pose.set_phi` and `Pose.set_psi`.

B.2 Data for assessing the stability effect of mutations in T4 Lysozyme protein

Predictions for the stability effect of mutations in the T4 Lysozyme protein were made using the PDB structure 2LZM [146], as the wild type. The PDB identifiers for structures of all the T4 Lysozyme variants used in our analyses (shown in Figs. 3.3 and 3.4), along with the reported $\Delta\Delta G$ and the pH values of the stability measurements are reported in Table B.1.

Experimental measurements of $\Delta\Delta G$ values in variants for which we do not have a matching protein structure (shown in Fig. 3.3 and 3.5) are taken from the dataset FireProtDB [131].

The AlphaFold prediction of the T4 Lysozyme protein structure, used in Fig. 3.3C and 3.5, was performed using the AlphaFold GoogleColab notebook [66].

Relaxation of mutant T4 Lysozyme structures for the *in silico* model in Figs. 3.3B,C and 3.4, was done using PyRosetta `FastRelax` with the score function `ref2015_cart`. Backbone flexible positions were restricted to the mutated residue and the neighboring residues in the sequence. Side chain flexible positions were restricted to the neighbors with alpha carbons within a 10 Å of the mutated residue.

B.3 Data for assessing the fitness effect of mutations in the RBD of SARS-CoV-2

Deep mutational scanning experiments from ref. [124] were used to obtain the SARS-CoV-2 RBD expression and binding to the ACE2 receptor. Predictions of the RBD-ACE2 binding strength, shown in Fig. 3.6B, were made using the co-crystallized protein structure with PDB identifier 6M0J [73]. The predictions for the RBD stability, shown in Fig. 3.6A, were made using the 6M0J structure with the ACE2 chain computationally removed. This computational removal was performed by selecting the RBD structure in PyMol [115] and exporting a pdb file from said selection.

variant	I3Y	I3P	V87M	D92N	R96H	R96N	R96D	R96W	R96Y	L99G	M102K	M102L	M106K	M120K	V149M	V149S	V149C	V149G	G156D	T157I
pdb	1L18	1L97	1CU3	1L55	1L34	3CDT	3C8Q	3FI5	3C80	1QUD	1L54	1L77	231L	232L	1CV6	1G06	1G07	1G0P	1L16	1L10
$\Delta\Delta G$	-2.3		-2.3	-1.4		-3.0	-3.5	-4.5	-4.7	-6.3	-6.9	-2.11	-3.4	-1.6	-2.8	-4.4	-2.0	-4.9	-2.3	-1.2
pH	6.5		5.4	5.7 -5.9		5.35	5.35	5.35	5.35	5.4	5.3	5.7	3	3	5.4	5.4	5.4	5.4	6.5	6
source	[84]	[33]	[43]	[94]	[145]	[89]	[89]	[89]	[89]	[157]	[30]	[59]	[79]	[79]	[43]	[159]	[159]	[159]	[49]	[50]

variant	I3V	T26S	S38N	S38D	G77A	A82P	A93T	M106I	M106L	E108V	T109N	T109D	N116D	S117F	M120Y	M120L	N144D	V149I	T151S	T152V
pdb	1L17	131L	1L61	1L19	1L23	1L24	129L	1P46	234L	1QUG	1L59	1L62	1L57	1TLA	1P6Y	233L	1L20	1G0Q	130L	1G0L
$\Delta\Delta G$	-0.4		0.6		0.4	0.8		0.2	0.5	0.7	0.1	0.6	0.6	1.1	-0.1	0.5		-0.1		0.2
pH	6.5	5.4	5.7 -5.9	5	6.5	6.5	5.4	5	3	5.4	5.7 -5.9	5.7 -5.9	5.7 -5.9	5.4	5	3	5	5.4	5.4	5.4
source	[84]	[102]	[94]	[95]	[85]	[85]	[102]	[90]	[79]	[157]	[94]	[94]	[94]	[6]	[90]	[79]	[95]	[159]	[102]	[159]

Table B.1: **Tables of all T4 mutants studied grouped by qualitative mutational effect** Each variant with a matching structure is listed along with the corresponding PDB entry. When available the pH and the $\Delta\Delta G$ from the experiment are provided.

Figure B.1: **Experimental RBD expression of SARS-CoV-2 in DMS experiments and the H-CNN predictions. (A)** Experimental expression of single-point mutation variants of SARS-CoV-2 RBD relative to the wild type (dots) as measured using yeast surface display experiments. Expression effect e is calculated as $\Delta \log_{10}(\text{MFI})$ where MFI is the mean fluorescent intensity. Mutations to G or P dominate the negative tail of expression effects. To better show the effect of all mutations, the negative portion of the colorbar is bounded by the expression of the most detrimental variant that is not a mutation to G or P. **(B)** The mean expression effect of mutations $\bar{e}$ is listed next to each site. **(C,D)** The H-CNN predicted stability effect of mutations (D), and the mean predicted stability effect for each site $\bar{E}$ (C), evaluated based on the computationally isolated RBD structure, are shown. Again the negative portion of the colorbar is bounded from below by the most destabilizing prediction not involving mutations to G or P.

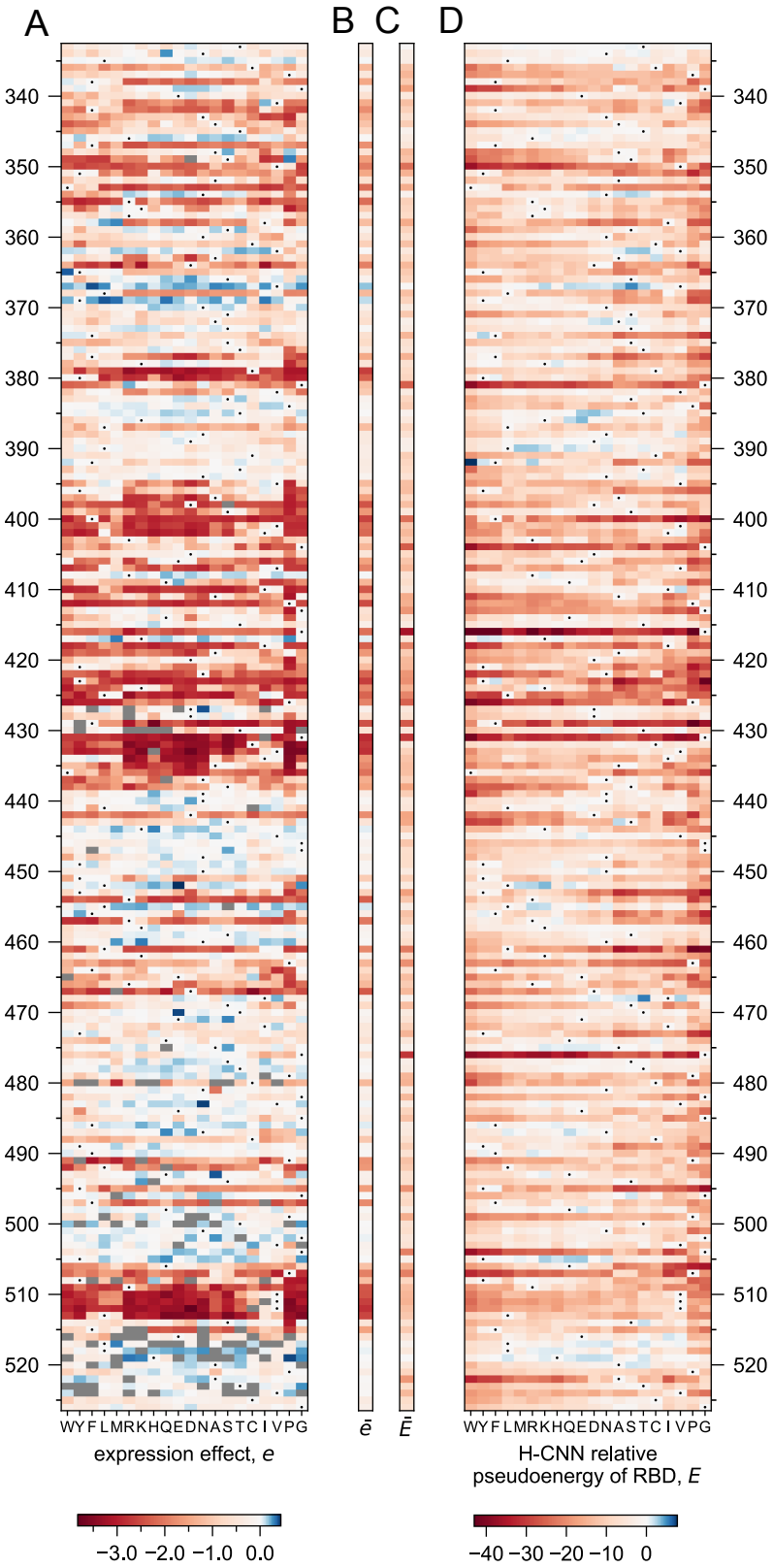
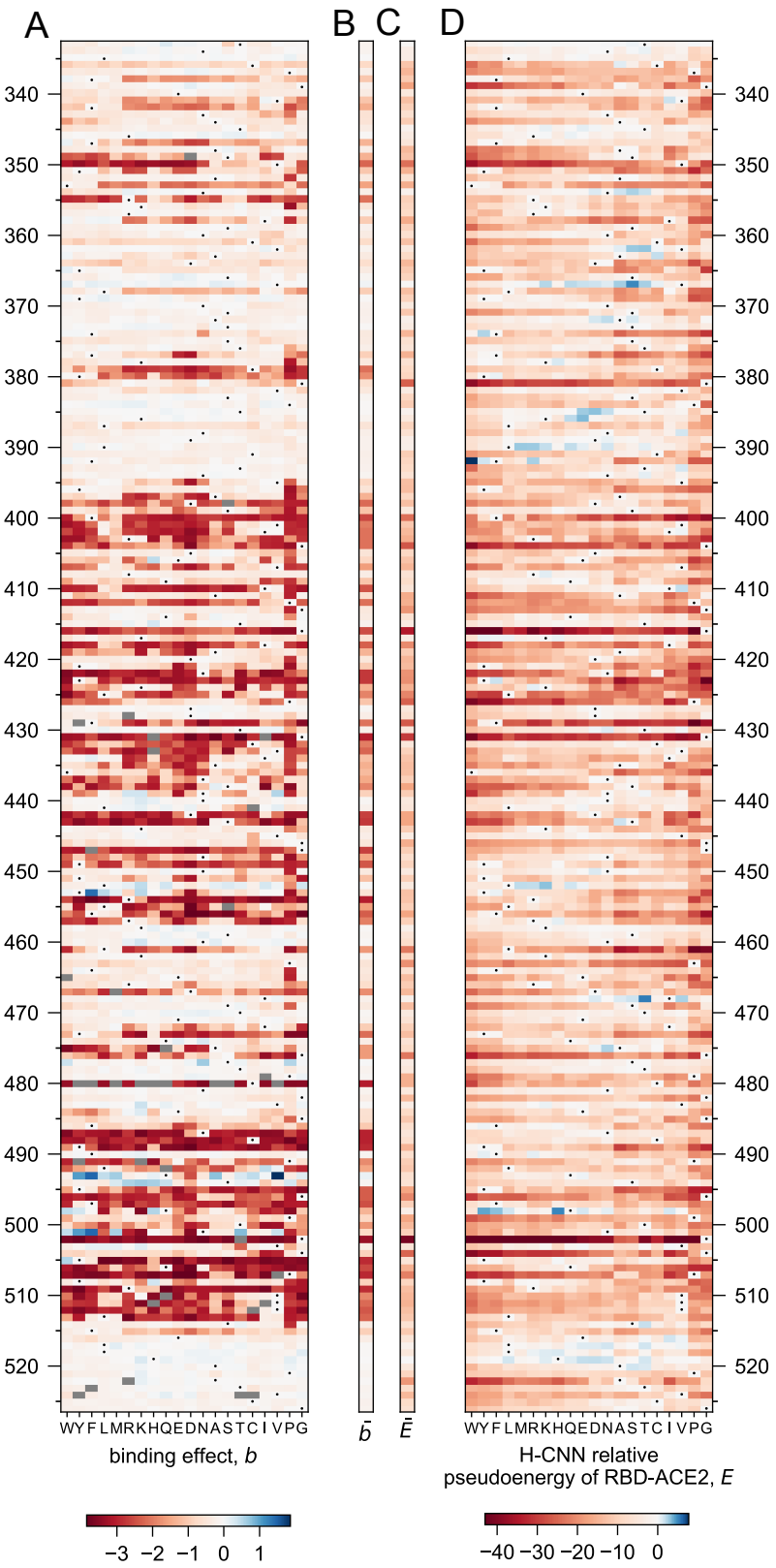

A
B
C
D
340
350
360
370
380
390
400
410
420
430
440
450
460
470
480
490
500
510
520
WY F L M R K H Q E D N A S T C I V P G
expression effect, e
$\bar{e}$
$\bar{E}$
H-CNN relative
pseudoenergy of RBD, E
−3.0 −2.0 −1.0 0.0
−40 −30 −20 −10 0

Figure B.2: **Experimental RBD-ACE2 binding affinity in DMS experiments and the H-CNN predictions. (A)** Experimental binding of single-point mutation variants of SARS-CoV-2 RBD to the human ACE2 receptor relative to the wild type (dots) as measured using yeast surface display experiments. Binding effect b is reported from $\Delta \log K_D$ where negative values represent weaker binding relative to the wild type. Mutations to G or P dominate the negative tail of expression effects. To better show the effect of all mutations, the negative portion of the colorbar is bounded by the expression of the most detrimental variant that is not a mutation to G or P. **(B)** The mean binding effect of mutations $\bar{b}$ is listed next to each site. **(C,D)** The H-CNN predicted effect of mutations (D) and the mean predicted stability effect for each site $\bar{E}$ (C), evaluated based on the crystal structure of RBD-ACE2 complex, are shown. Again the negative portion of the colorbar is bounded from below by the most destabilizing prediction not involving mutations to G or P.


A
B
C
D
340
350
360
370
380
390
400
410
420
430
440
450
460
470
480
490
500
510
520
WYFLMRKHQEDNASTCIVPG
binding effect, b
$\bar{b}$
$\bar{E}$
H-CNN relative
pseudoenergy of RBD-ACE2, E
−3 −2 −1 0 1
−40 −30 −20 −10 0