2022

Structure-Based Drug Discovery and Development of Protein Structure Prediction Tools Using an Empirical Force Field

Noah B. Herrington
Virginia Commonwealth University

Structure-Based Drug Discovery and Development of Protein Structure Prediction Tools Using an Empirical Force Field

A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy at Virginia Commonwealth University

By

Noah Benjamin Herrington

Bachelor of Science, Randolph-Macon College, 2018

Advisor:

Glen E. Kellogg, Ph.D.

Professor, Department of Medicinal Chemistry

Virginia Commonwealth University

Richmond, Virginia

August 2022

Acknowledgements

I would like to begin by first thanking the Virginia Commonwealth University School of Pharmacy, Graduate School, and Department of Medicinal Chemistry for the amazing opportunity of entering this program that was presented to me. I have been truly fortunate to conduct my studies with generous financial support, without which this endeavor would not have been possible. Further, I am thankful to my committee members and lab mates, especially Claudio Catalano and Mohammed AL Mughram, for sharing their expertise and advice while I was first learning basic computational chemistry.

I would next like to thank my adviser, Dr. Glen E. Kellogg, who first recruited me to his lab as his latest padawan after I showed interest in moving away from synthetic organic chemistry and into computational chemistry. Working in Dr. Kellogg's lab opened my eyes to the possibilities for studying biological events and drug design *in silico*. He not only encouraged me to think creatively and scientifically and try different techniques in the lab, branch out and learn new techniques, such as basic programming, but he also offered a tremendous amount of support and wisdom when I lost my sense of direction and purpose in my graduate school career. This included a few well-deserved reality checks. His support and perspective have shaped the scientist I have become.

Finally, I would love to thank my family and friends for their love and support that gave me the resilience I needed to stay with this program. My mother, brother, and pets offered unconditional love I often took for granted, and my friends reminded me that the greatest challenges of our program were not insurmountable. If I have not said so enough already, my genuine thanks goes out to every single one of you.

"Ludwig Boltzmann, who spent much of his life studying statistical mechanics, died in 1906, by his own hand. Paul Ehrenfest, carrying on the work, died similarly in 1933. Now it is our turn to study statistical mechanics." – David L. Goodstein, States of Matter (1985)

Table of Contents

Table of Figures

Table of Tables

Abstract

STRUCTURE-BASED DRUG DISCOVERY AND DEVELOPMENT OF PROTEIN STRUCTURE PREDICTION TOOLS USING AN EMPIRICAL FORCE FIELD

By Noah B. Herrington

A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy at Virginia Commonwealth University

Virginia Commonwealth University, 2022

Major Director:

Glen E. Kellogg, Ph.D.

Professor, Department of Medicinal Chemistry

Traditional drug discovery has rapidly accelerated thanks to development of computational molecular modeling. The crucial component that these computational studies hinge upon is having a well-defined, and energetically favorable structure. Structures of proteins and ligands that meet these criteria are important for accurately simulating models used to study drug binding. To demonstrate the role of accurate structure simulation in the study of these events, this thesis presents, first, a story examining the problem of accurate structure modeling of ionizable residues within protein structures, specifically aspartic acid, glutamic acid, and histidine. I present our method, which uses the HINT force field to simulate "titration" of these residues and study which hydropathic environments may contribute to stabilization of certain protonation states. We further use this data to construct and cluster together pH-tunable hydropathic interaction maps, detailing the kinds of interactions these residues make with their environments in low and high-pH situations.

The second story describes identification of new, potent inhibitors against eIF4A1 (eukaryotic initiation factor 4A1), driven by computational techniques. This story describes a pharmacophoric virtual screen of chemical databases for novel inhibitors, based on the structure of Rocaglamide A (RocA), a natural product inhibitor of eIF4A1. After docking and HINT scoring studies of hit compounds, we identified many highly potent compounds. Computational studies have yielded a reasonable binding mode for this series of compounds and suggest design of new, more potent compounds with better drug-like properties.

The final story builds upon our compilation of hydropathic interaction maps in the design of a protein-protein interface optimization program that will be the roots of a protein-protein docking tool. We compile vast amounts of hydropathic map data, detailing what we call residue "hydropathic valences," for this purpose. The tool implements a genetic algorithm for population-weighted choice of map combinations for residues at a protein-protein interface. Our model is currently being trained on publicly available, high-resolution crystal structures. We hope for development of this tool to be the beginning of returns made on a long series of chapters of data collection for this purpose.

This thesis is a record of diligent efforts to apply HINT to novel drug discovery and protein structure prediction tools. It will demonstrate the integral role of using or creating accurate structure models for studying protein structure and how these studies may ultimately be used for development of new clinical therapeutics. Let this work also stand as a testament to the power of computational techniques to efficiently simulate real-world biomolecular events on an atomic scale in a way that even allows this translation from *in silico* theory to potentially *in vivo* reality. Let it be astounding to the reader, as it was for me.

Chapter 1: Introduction[†]

Every natural process, no matter how small, has an explanation based on the smallest building blocks of matter. Perhaps, Richard Feynman said it best: "…all things are made of atoms, and everything that living things do can be understood in terms of the jigglings and wigglings of atoms."[2] One thing that living things can do is bind drug molecules, which has the same basis in erratic movement of atoms. The binding of drug molecules to specific proteins has much to do with structural complementarity between the ligand and protein. That is, to say, the structural components of the ligand molecule, including its hydrophobic or electrostatic character, connectivity, and shape, directly affect its ability to non-covalently bind to a protein, whose structure bears a cavity amenable to the ligand. This is the basis for the so-called 'lock-and-key' mechanism for protein-ligand binding, proposed by Emil Fischer in 1894.[3] Another model for understanding this phenomenon is known as 'conformational selection.' This model accounts for the dynamic nature of protein movement and conformational changes and poses that binding of a ligand 'chooses' the most optimal protein conformation for binding together both species. Many current drug discovery projects intend to exploit this structural complementarity relationship to design new and better drug therapies because higher complementarity generally elicits strengthened therapeutic activity. This deeper relationship between protein and ligand structure is crucial for understanding how we may further progress our drug discovery efforts using computational techniques.

[†] This chapter contains sections that have been adapted from Herrington, N. B.; Kellogg, G. E. **2021**[1]

It is well-known that structure determines function with respect to proteins. For this reason, the past few decades have seen increased efforts to elucidate protein structures with hopes of learning and understanding how their functions are related to their structures. It is important to be mindful that the primary amino acid sequence has a strong impact on the manner in which proteins fold into higher-order secondary, tertiary, and quaternary structures.[4] The unique identities of amino acids in this chain can be integral to forming intramolecular interactions stabilizing the overall structure of the complete protein. The shape and surface residues of a particular protein ultimately determine what other species it interacts with and its function. Proteins containing similar strings of amino acids often adopt similar three-dimensional structures and similar roles within a cell. These proteins belong to what is called the same *protein family*. It is this structure-function relationship that is at the heart of modern drug discovery because certain aspects of protein structures, possibly specific residues or secondary structures, are crucial for maintaining their functions, making them viable targets for small-molecule therapies. Small molecules, in this instance, can act as antagonists (i.e., inhibitors), blocking interactions of the protein with endogenous substrates or other signaling proteins, or as agonists (i.e., activators) by stabilizing proteins in a specifically active conformation. This understanding of protein structure is a core component of computational drug discovery efforts.

Computer-Aided Drug Design

Computer-Aided Drug Design (CADD) has demonstrated itself to be a valuable asset in drug discovery campaigns and has seen a greater presence in the limelight over the past few decades. Some of the early interest in this area, perhaps, was marked by *Fortune's* 1981

publication of "Next Industrial Revolution: Designing Drugs by Computer at Merck."[5] Around this time, high-throughput screening (HTS) strategies were gaining popularity as a way of rapidly screening multitudes of compounds, but many projects were limited by the costs of this technique. CADD methods, like virtual screening (VS), soon emerged as an alternative to HTS, offering another fast method for screening vast quantities of compounds, but with the predictive power of filtering out compounds unlikely to bind protein targets of interest and therefore not elicit the desired therapeutic activity. This allowed researchers to preliminarily screen compounds before testing them in the lab, saving much needed time and money.

The most important CADD technique is possibly molecular docking. This technique is designed to identify and exploit structural complementarity between a drug molecule and a protein and make an educated prediction of the most energetically favorable pose of the molecule once bound to the protein, based on hydropathic interactions made between both molecules. For this reason, it is also used as an alternative strategy to crystallographic techniques intended to elucidate the structure of a protein-ligand complex. Docking techniques can be an integral component of many other techniques, including VS, molecular dynamics (MD), and 3D Quantitative Structure-Activity Relationship (3DQSAR) studies. The most powerful product of docking is a working model of a ligand binding pose within an active site, which can be used to potentially explain various aspects of activity from protein-ligand binding, as well as a starting place for designing new, possibly improved therapeutics. Additionally, a docked model may become the basis for designing new laboratory experiments that either confirm or reject hypotheses for structural bases for activity. Different algorithms exist for molecular docking, including those for rigid-body docking and

flexible docking. Rigid-body docking most closely replicates the 'lock-and-key' mechanism of ligand binding by limiting the conformational flexibility of the binding site residues. Flexible docking, obviously, is more liberal in terms of the binding site flexibility and more closely imitates the 'conformational selection' mechanism of ligand binding.[3] Both have utility for modeling ligand binding, where rigid body docking is particularly useful for docking of ligands based on poses of known reference ligands and for being the most expeditious of the two methods, while flexible docking may simulate completely new interactions that may be unexplored for new/unknown binders. Examples of the use of docking will be showcased later in this thesis for its application to the discovery of novel anti-cancer therapeutics.

In addition to the protein residues and small molecules of a structure, it is also important to consider the role of solvent when conducting CADD. Water possesses both hydrogen bond acceptor and donor components that can play their own parts in ligand binding. It is often the case that water participates as a cofactor facilitating the binding of a small molecule. Below, the role of water, solvation, and ligand solubility will be discussed in further detail, as these concepts are foundational to many of our molecular modeling studies.

The Hydrophobic Effect and the HINT Force Field

Water is widely regarded to be a substance integral to the survival of all organisms. The human body is roughly composed of 70% water and plays a role in numerous physiological and cellular processes. Ironically, in spite of its highly polar nature, it is responsible for what is known as the 'hydrophobic effect.'[6-8] Fundamentally, this concept is characterized by the clustering and compaction of hydrophobic species together in the presence of water to minimize contact with the hydrophilic solvent. This results in a number

of phenomena, including the separation of oil and water and, more importantly, specific folding of protein structures. The hydrophobic effect, in this case, is twofold in that it influences the packing of hydrophobic residues into the protein core and the stabilization of protein three-dimensional structure by forming a solvent network with the polar residues of the protein exterior. Evidence of this secondary effect can be seen in estimates of between 10% and 20% higher solvent density in the first layer of solvent surrounding proteins than in bulk water.[6] In the context of drug discovery, the focus is on solvation of a protein and a protein-ligand complex. It is important to note that either solvation process is also a thermodynamic process, whose free energy can be calculated, according to:

$$\Delta G = \Delta H - T\Delta S,$$

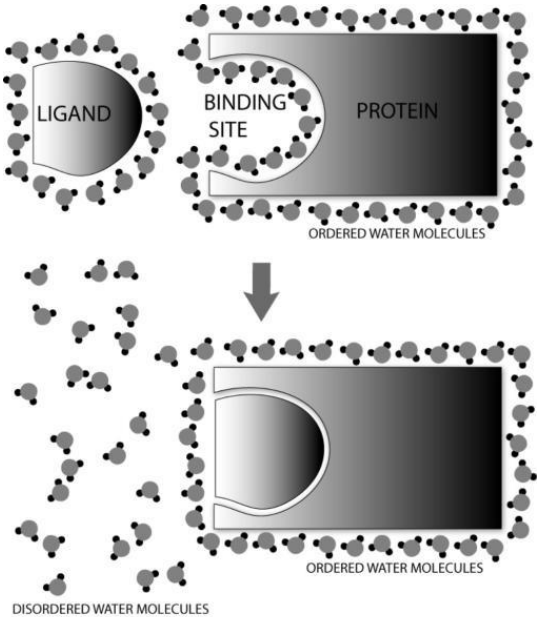


Figure 1.1. Ligand displacement of water molecules bound to a protein site. In this representation, a ligand fills a protein binding site and displaces an array of bound water molecules, ejecting them into bulk solvent. This, overall, contributes to a higher level of entropy from the perspective of water molecules.

The significance of this equation is that solvation events have both enthalpic and entropic components, represented by the ΔH and ΔS terms, respectively. Roughly speaking, the enthalpic term largely refers to water molecules' ability to form hydrogen bonds with themselves, a protein surface, and/or a bound ligand molecule, while the entropic term refers to the amount of 'disorder' among the water molecules. In general, stronger hydrogen bonds and lower enthalpy obviously contribute to a lower and more favorable Gibbs energy and state in this binding equilibrium, while

increased entropy also gives a lower ΔG. The trouble is that hydrogen bonds are best formed when there is less disorder to disrupt them, so enthalpy and entropy often work against each other. This is referred to as 'enthalpy-entropy compensation.' [7-9] In terms of ligand binding events, a ligand must disrupt the hydrogen bonding network of waters within its binding pocket and displace them before forming its own interactions with the protein, which elevates the entropy of waters in bulk solvent (Figure 1.1).[10] Therefore, in order for this binding event to be spontaneous, the free energy of the bound ligand, including the enthalpic contributions from formed hydrogen bonds and the increased disorder of surrounding solvent, must be lower and more favorable than the solvated pocket. The hydrophobic effect is therefore highly complex and multifaceted, but it is important for understanding how solvation affects stabilization of the drug-bound state of a protein target.

In the event of drug molecule binding, there are largely two types of water molecules that can be displaced, according to Spyrakis et al.:[9] 1) waters in large cavities eventually occupied by the bulk of a ligand structure and 2) waters displaced from addition of substituents to the ligand structure. Waters in the first situation are easily displaced, as they are often found in hydrophobic sites, since these are the most cavernous regions within protein structure, but waters in situation two are often trapped in hydrophilic regions and are more difficult to remove. In both cases, drug design must consider the change in free energy between the unbound and bound states. Transitioning from the unbound state to the bound state already implies an increase in enthalpy from water's perspective, due to disruption of any water-based hydrogen bonds within its occupied pocket. This, of course, though, is accompanied by a favorable increase in entropy, as more water is freed from the binding site and is allowed to enter its bulk solvent body to hydrogen bond with itself.

Therefore, the free energy of a ligand binding must be more favorable than that of the solvated state. Unfortunately, many modern force fields in CADD are designed to optimize the enthalpic effects for ligand binding, rightfully so, as the enthalpic term of free energy is often the greatest contributor and easiest to simulate. However, the remaining entropic term must not be left out when considering the favorability of ligand binding.

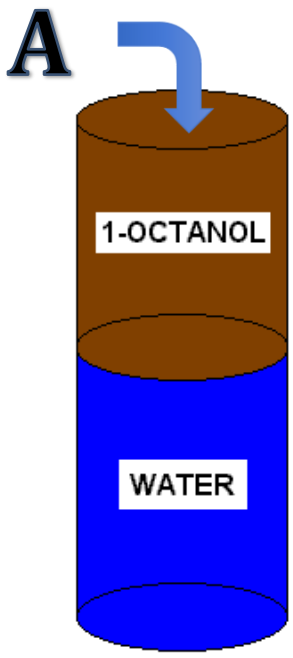


Figure 1.2. Partition experiment for a compound A between layers of 1-octanol and water. The logarithm of the ratio of the compound's concentration in each layer is used to assess the compound's solubility, an important factor to consider for drug discovery.

Even considering the complicated nature of hydrophobicity, a great deal of information can be extracted from solubility and solvation events. Specifically, solubility is an important factor to consider for drug design because it affects a ligand molecule's ability to enter solution, be absorbed into the bloodstream, and pass through cell membranes. For this reason, the partition coefficient of a drug in a mixture of 1-octanol and water ($LogP_{o/w}$) is often measured to assess a drug candidate's solubility (Figure 1.2). Solvation events have dual enthalpic and entropic nature that, therefore, must be components of $LogP_{o/w}$. In fact, the ratio of a compound's partition into 1-octanol versus water can be treated as an equilibrium constant, containing important information related to the favorability of it interacting with either layer. This fact is the basis for the design of HINT[11,12] (Hydropathic INTeractions) force field in our lab. In short, HINT considers experimental partial $LogP_{o/w}$ values for atomic components of molecular fragments as a manner of

estimating complete calculated partition ($cLogP_{o/w}$) for participants in intermolecular interactions and the free energy of their binding together. HINT calculates atomic "hydrophobic atom constants," originally proposed by Abraham and Leo,[13] derived from "fragment constants" calculated from experimental partition data.[14] It further calculates an interaction score between two atoms *i* and *j* in space, according to the equation:[9]

$$H_{Total} = \sum_i \sum_j b_{ij} = a_i a_j S_i S_j T_{ij} R_{ij} + 50 r_{ij}$$

Here, *a* is the hydrophobic atom constant, *S* is the solvent-accessible surface area of the atom, T_{ij} is a logic function indicating the favorability of the interaction as -1 or +1, R_{ij} is exponential function of the distance between atoms *i* and *j*, and r_{ij} is an implementation of the Lennard-Jones potential. In general, 515 HINT units ≈ 1 kcal mol^{-1} and is thus able to estimate the free energy of an interaction.[15,16] HINT is a core component of all studies conducted and showcased here.

Computational Methods for Predicting Protein Structure

Today, there are over 190,000 structures of proteins deposited into the PDB, the growth of which has been largely outpaced by the almost 232,000,000 sequences deposited into UniProt.[17] The structures in the PDB represent only nearly a tenth of a percent of the sequences known for proteins. This illustrates the scale of the protein structure prediction problem: where structure elucidation methods unfortunately fail to produce protein structural data, computational techniques can hopefully bridge the gap and provide new methods for obtaining this information. Cyrus Levinthal, if he was alive today, might have warned many of us, whose careers have been built on studying protein structure with hopes

of one day predicting it. The fact that a completely random search for the native structure of a protein, including the correct conformations of residue backbones and side chains, could take a seemingly endless amount of time, but that it takes seconds or less for it to fold in Nature has been termed the "Levinthal Paradox." In spite of this (or in defiance of it), a multitude of different methods for predicting protein structure already exist. Additionally, the scientific community's processing power has rapidly increased since Levinthal's proposal of this problem. To illustrate, revolutionary technology from 1964 was developed in the form of a new disk drive that could store approximately 7.6 megabytes (MB) of data. Today, most email servers can send a message with an attachment of up to 25 MB, and our fastest supercomputers can compute quadrillions of calculations per second.[18] That said, "protein structure prediction" is an extremely broad term covering a multitude of different techniques with various purposes ranging from the smallest, most specific to the largest, most complex parts of proteins. According to Anfinsen's Dogma,[19] the only necessary information for a protein to properly fold is its primary amino acid sequence. Modern protein structure prediction methods are designed around this maxim.

Tools that perform protein structure prediction generally fall into one of two categories: template-based and *ab-initio* structure prediction.[20] Template-based, i.e., homology modeling methods rely on existing structural data from known sequences as input for modeling similar sequences. The theory behind this practice is that similar sequences are likely to have similar structures. When two sequences are similar enough (exceeding 80% identity or similarity), homology modeling can be an accessible and effective method for modeling structures that might otherwise be unattainable by crystallographic or other

acquisition methods. However, these template-based methods become much less effective when the pair's sequence identity falls below 30%.[21]

Ab-initio methods, perhaps, are a misnomer, as they are implied to be entirely "template-free." However, all protein structure prediction methods rely on sequence-structure data of some nature. For example, one of the most popular protein structure modeling tools is Rosetta,[22,23] developed by David Baker's lab. The premise of Rosetta is to take micro-sequences of length 3-9 residues from the Protein Data Bank (PDB)[24] to progressively build full-length structures, but even this is largely "template-based," though the templates are much smaller than full-sized proteins. Another *ab initio* has been developed recently in the form of AlphaFold,[25,26] which has been heralded by many as the "solution" to the protein folding problem.[27,28] Its artificial intelligence algorithm is designed to predict the structure of a protein, given its sequence, based on what its training on protein structural features from structures imported from the PDB. This effectively ensures that it, too, is not independent of sequence-structure data for model prediction. For the purposes of this thesis, the categories of template-based and *ab-initio* methods shall still remain separate.

Considering the effect that AlphaFold has had on the scientific community, it is important to address some of its shortcomings. Its arrival has certainly been met with a mixture of praise and criticism. Many computational and structural biologists have found it incredibly useful for modeling structures of interesting proteins for virtual screening[29,30] and protein-protein complexes.[31,32] However, AlphaFold is not without its shortcomings.[33] For one, it does not compute structures of partner proteins as part of multimers and, therefore, does not compute well the positions of interface residues, which are then free to drift away

from folded residues into space. It also fails where little sequence-structure data is available for alignment (another reason AlphaFold is highly similar to template-based modeling methods). Additionally, it lacks the ability to predict positions of ligands, such as metal ions or cofactors, which may be integral components for determining the folding of a protein. Most notably, for purposes of this thesis, AlphaFold makes no effort to simulate the fitness of variations in protonation states of ionizable residues. Structural biology is full of examples of protonated ionizable residues contributing to the structure and function of enzymatic action.[34-36] Our methods for protein structure prediction are evolving with each year; with this evolution, the best of our methods will hopefully learn to consider these important features of proteins when predicting their folds and functions.

Computational Studies of pH, pK_a, and Protonation States

One important aspect of the relationship between protein structure and function is the dependence of protein structure on pH and protonation states of constituent residues. Histidine (HIS), for example, has a nominal pK_a of 6.00,[37] situated closely enough to physiological pH that its imidazole sidechain can act either as a cationic dual hydrogen bond donor or a neutral donor and acceptor depending on its local pH environment. That is, the resultant influence of a residue's neighborhood, comprised of the hydrogen bond donors, acceptors, charged species, etc. that influence the solution pH surrounding it.[38] The importance of histidine's protonation state in the so-called "catalytic triad" of serine, histidine, and aspartate in serine proteases was shown decades ago for trypsin.[39,40] The pH-dependence of protein function is a well-established principle and has promoted extensive research into identifying optimum pH for activity of various other macromolecules.[41]

The pK_as of aspartic acid (ASP) and glutamic acid (GLU) when isolated or in model peptides are reported to be 3.65 and 4.25, respectively,[37] making them functionally similar residues and leaving them both largely deprotonated at physiological pH. These pK_as are not static, and large deviations from these values are not uncommon. For example, the active site of bacteriorhodopsin contains an aspartic acid with an experimental pK_a of 7.68.[42]

Unfortunately, protein structure elucidation by X-ray crystallography or cryogenic electron microscopy are seldom of sufficient resolution to determine locations of hydrogens, due to their extremely low electron density. X-ray crystallography detects protons only under difficult-to-achieve conditions such as resolution ~1 Å.[43] Such resolution is not yet possible with cryo-EM. While neutron diffraction experiments can overcome this problem,[44,45] as it is detecting nuclei rather than electrons, experimental constraints, such as required crystal sizes, availability of neutron sources, and others, make neutron diffraction-derived structures for proteins quite rare. Multidimensional nuclear magnetic resonance methods can be applied to protein structure determination,[46] but only under certain conditions like protein size and solubility. Because NMR directly probes hydrogens, it can be used for pK_a determination of specific residues,[47,48] but this is only a probe of the residue under the NMR experimental conditions, which may differ greatly from its native physiological or solution conditions. In general, it is quite difficult to discern structural reasons for residue pK_a shifts experimentally, although this is a quite active area of computational research as many reports have been published suggesting what types of environments stabilize shifts.[49-51] Interestingly, experimental methodologies such as NMR perform well in determining pK_as for surface ionizable residues but are less applicable to buried residues.[52]

Much of the effort to study protonation of ionizable residues via computational means has focused on predicting their pK_as by understanding the effects of other residues in the local environment. Li et al. developed a method, known as PROPKA, to empirically calculate pK_a values impacted by nearby residues.[53] In this model, hydrogen bonding to aspartates and glutamates stabilizes their deprotonated forms and lowers their pK_as. Spassov and Yan[54] utilized CHARMM[55] to develop a molecular dynamics-based approach to predict pK_a values of titratable groups. Several factors of 3D protein structure determination—and the resulting structural model—can compromise such predictions, e.g., uncertainties in sidechain conformations if the collected data resolution is too low.[56] It should be emphasized that model building is an important aspect of crystallography, which is collection of experimental data, so even presentation and use of crystallographic data carries some uncertainty. This highlights the importance of collecting, refining, and beginning any computational study with accurate structural data, including (perhaps, especially) modeling of hydrogen positions.

Three-Dimensional Interaction Homology

Since the dawn of protein structure elucidation, our understanding of the roles and contributions of interatomic interactions between protein residues toward biomolecular structural organization has evolved dramatically. Each of the 20 amino acid residues, regardless of how many unique protein structures they compose, is likely to situate itself within a limited set of environments with a unique system of interactions of varying magnitude, type, and loci. Our model describes four classes of interactions: favorable polar (e.g., hydrogen bond, acid-base), unfavorable polar (acid-acid, base-base, repulsive

Coulombic), favorable hydrophobic (hydrophobic-hydrophobic, hydrophobic packing, π-π stacking) and unfavorable hydrophobic (hydrophobic-polar, desolvation).

Importantly, interactions with the environment of each constituent residue of a protein contributes in some part toward its rotameric structure and the protein's overall secondary, tertiary, and quaternary structure. Our hypothesis is that each residue has a "hydropathic valence" that must somehow be satiated by nearby interacting groups. Hydrophobic residues such as phenylalanine and leucine, by interacting with other hydrophobic groups, pack together to avoid water, while polar residues, such as the three of this study, favor environments where they can engage in polar interactions, e.g., hydrogen bonding, with other residues or water. Thus, obviously, 3D protein structure is not driven by "primary" structure, but by the hydropathic interactions that each residue must make based on its type and sidechain and backbone conformations.

In our first report to address this concept, we calculated 3D hydropathic interaction maps to visualize and probe all possible environments of tyrosine (TYR) using a dataset of ~30,000 residues. Our analysis organized all of our TYR residues into 262 unique, backbone-dependent environments, each with a unique map encoding the specific interactions made by the residue in that environment.[57] A similar analysis with over 57,000 alanine (ALA) residues, separately calculating backbone-environment and sidechain-environment maps, yielded 136 and 150 backbone- and sidechain-dependent maps, respectively, despite ALA's simplicity. We concluded that ALA's mapped environments are a new and insightful form of structural motif.[58] Recently, in our report on phenylalanine, tryptophan, and tyrosine, we showed that the subtle effects of π-π and π-cation interactions are encoded in their 3D

hydropathic interaction maps.[59] In a report on serine and cysteine we highlight the major structural features—similarities and differences—between these two isosteric residues.[60] Importantly, our analyses describe residues by cataloguing their environments in terms of interactions and not identity. A water molecule oriented for a residue can play the same "acidic" role as a TYR–OH or a LYS–NH_3^+ to satisfy its hydropathic valence. Protein structure is driven by the set of these hydropathic interactions for each residue.

Additionally, the kinds of hydropathic interactions needed to fulfill each residue's hydropathic valence can certainly be exploited for drug discovery efforts. The hydropathic valences we have identified for each residue type are the same for interactions with other residues as they are for interactions with incoming small molecule ligands, meaning that docking of small molecules utilizes essentially the same information to predict the most favorable binding poses for ligands. This principle is important to understand how molecular docking functions in a later chapter of this thesis.

The importance of accurate structure modeling is the primary theme of this thesis, which is at the heart of three different studies, described below. In Chapter One, I present our method for optimizing the ionization states of aspartic acid, glutamic acid, and histidine and mapping their hydropathic environments. We conclude that these environments, secondary structure, and solvent accessibility are tied together, knowledge of which can be indispensable for developing protein structure prediction tools. Chapter Two examines the pharmacophoric virtual screen of the Rocaglamide A (RocA) binding site of eukaryotic initiation factor A1 (eIF4A1), resulting in identification of highly potent inhibitors. Our modeling studies suggest a potential binding mode for our compounds, which will be used

for design of new, hopefully more effective and more drug-like compounds. Finally, Chapter Three explores the development of a protein-protein interface optimization tool and precursor to a protein-protein docking tool. It implements a genetic algorithm that compiles the vast hydropathic environment map data at our disposal, which detail the most favorable environments surrounding all residue types, and performs operations, including population-weighted selection of maps for interface residues, crossover and mutation algorithms to simulate Darwinian natural selection, and ultimately building the most favorable model. Currently, our model suggests our scoring function may correlate with RMSD from a parent crystal structure. While the results of these studies speak for themselves, to an extent, it should be emphasized that their success was dependent on taking advantage of quality structural data. To reiterate, the availability of protein structural data opens many possibilities for drug discovery, a glimpse of which is presented here.

References

1. Herrington, N. B.; Kellogg, G. E. 3D Interaction Homology: Computational Titration of Aspartic Acid, Glutamic Acid, and Histidine Can Create pH-Tunable Hydropathic Environment Maps. *Front. Mol. Biosci.* **2021**, *8*, 773385.
2. The Relation of Physics to Other Sciences. https://www.feynmanlectures.caltech.edu/I_03.html (accessed April 25, 2022).
3. Mechanism of Enzyme Action. http://chemistry.elmhurst.edu/vchembook/571lockkey.html (accessed April 25, 2022).

4. Protein Structure. https://www.nature.com/scitable/topicpage/protein-structure-14122136/ (accessed April 27, 2022).
5. Sliwoski, G.; Kothiwale, S.; Meiler, J.; Lowe, Jr., E. W. Computational Methods in Drug Discovery. *Pharmacol. Rev.* **2014**, *66*, 334-395.
6. Chandler, D. Interfaces and the Driving Force of Hydrophobic Assembly. *Nature*, **2005**, *437*, 640-647.
7. Levy, Y.; Onuchic, J. N. Water Mediation in Protein Folding and Molecular Recognition. *Ann. Rev. Biophys. Biomol. Struct.* **2006**, *35*, 389-415.
8. Schauperl, M.; Podewitz, M.; Waldner, B. J.; Liedl, K. R. Enthalpic and Entropic Contributions to Hydrophobicity. *J. Chem. Theory Comput.* **2016**, *12*, 4600-4610.
9. Spyrakis, F.; Ahmed, M. H. Bayden. A. S.; Cozzini, P.; Mozzarelli, A.; Kellogg, G. E. The Roles of Water in the Protein Matrix: A Largely Untapped Resource for Drug Discovery. *J. Med. Chem.* **2017**, *60*, 6781-6827.
10. Sarkar, A.; Kellogg, G. E. Hydrophobicity—Shake Flasks, Protein Folding and Drug Discovery. *Curr. Top. Med. Chem.* **2010**, *10*, 67-83.
11. Kellogg, G. E.; Semus, S. F.; Abraham, D. J. HINT: A New Method of Empirical Hydrophobic Field Calculation for CoMFA. *J. Comput. Aided Mol. Des.* **1991**, *5*, 545-552.
12. Kellogg, G. E.; Abraham, D. J. Hydrophobicity: Is LogP(o/w) More Than the Sum of its Parts? *Eur. J. Med. Chem.* **2000**, *35*, 651-661.
13. Abraham, D. J.; Leo, A. J. Extension of the Fragment Method to Calculate Amino Acid Zwitterion and Side Chain Partition Coefficients. *Proteins: Struct., Funct., Genet.*, **1987**, *2*, 130-152.

14. Hansch, C.; Leo, A. Substituent Constants for Correlation Analysis in Chemistry and Biology. *J. Wiley and Sons, New York*, **1979**.
15. Cozzini, P.; Fornabaio, M.; Marabotti, A.; Abraham, D. J.; Kellogg, G. E.; Mozzarelli, A. Free Energy of Ligand Binding to Protein: Evaluation of the Contribution of Water Molecules by Computational Methods. *Curr. Med. Chem*. 2004, *11*, 3093−3118.
16. Burnett, J. C.; Botti, P.; Abraham, D. J.; Kellogg, G. E. Computationally Accessible Method for Estimating Free Energy Changes Resulting from Site-Specific Mutations of Biomolecules: Systematic Model Building and Structural/Hydropathic Analysis of Deoxy and Oxy Hemoglobins. *Proteins: Struct., Funct., Genet.* **2001**, *42*, 355−377.
17. The Uniprot Consortium. UniProt: The Universal Protein Knowledgebase in 2021. *Nucleic Acids Res.* **2021**, *49*, D480-D489.
18. Then and Now: Computing Power. National Nuclear Security Administration. https://www.energy.gov/nnsa/articles/then-and-now-computing-power (accessed June 17, 2022).
19. Anfinsen, C. B. Principles That Govern The Folding of Protein Chains. *Science*, **1973**, *181*, 223-230.
20. Deng, H.; Jia, Y.; Zhang, Y. Protein Structure Prediction. *Int. J. Mod. Phys. B.* **2018**, *32*, 1840009.
21. Xieng, Z. Advances in Homology Protein Structure Modeling. *Curr. Protein Pept. Sci.* **2006**, *7*, 217-227.
22. Simons, K. T.; Kooperberg, C.; Huang, E.; Baker, D. Assembly of Protein Tertiary Structures from Fragments with Similar Local Sequences Using Simulated Annealing and Bayesian Scoring Functions. *J. Mol. Biol.* **1997**, *268*, 209-225.

23. Rohl, C. A.; Strauss, C. E. M. Misura, K. M. S.; Baker, D. Protein Structure Prediction Using Rosetta. *Methods Enzymol*. **2004**, *383*, 66-93.
24. Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, B. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235-242.
25. Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A.; Bridgland, A.; Meyer, C.; Kohl, S. A. A.; Ballard, A. J.; Cowie, A.; Romera-Paredes, B.; Nikolov, S.; Jain, R.; Adler, J.; Back, T.; Petersen, S.; Reiman, D.; Clancy, E.; Zielinksi, M.; Steinegger, M.; Pacholska, M.; Berghammer, T.; Bodenstein, S.; Silver, D.; Vinyals, O.; Senior, A. W.; Kavukcuoglu, K.; Kohli, P.; Hassabis, D. Highly Accurate Protein Structure Prediction with AlphaFold. *Nature.* **2021**, *596*, 583-389.
26. Lupas, A. N.; Pereira, J.; Alva, V.; Merino, F.; Coles, M.; Hartmann, M. D. The Breakthrough in Protein Structure Prediction. *Biochem. J.* **2021**, *478*, 1885-1890.
27. Callaway, E. What's Next for the AI Protein-Folding Revolution. *Nature*, **2022**, *604*, 234-238.
28. Toews, R. AlphaFold is the Most Important Achievement in AI – Ever. *Forbes*, **2021**. https://www.forbes.com/sites/robtoews/2021/10/03/alphafold-is-the-most-important-achievement-in-ai-ever/?sh=210529886e0a (accessed June 6, 2022).
29. Weng, Y.; Pan, C.; Shen, Z.; Chen, S.; Xu, L.; Dong, X.; Chen, J. Identification of Potential WSB1 Inhibitors by AlphaFold Modeling, Virtual Screening, and Molecular Dynamics Simulation Studies. *Evid. Based. Complement. Alternat. Med.* **2022**, *2022*, 4629392.
30. Yang, C.; Alam, A.; Alhumaydi, F. A.; Khan, M. S.; Alsagaby, S. A.; Al Abdulmonem, W.; Hassan, Md. Imtaiyaz; Shamsi, A.; Bano, B.; Yadav, D. K. Bioactive Phytoconstituents

as Potent Inhibitors of Tyrosine-Protein Kinase Yes (YES1): Implications in Anticancer Therapeutics. *Molecules*, **2022**, *27*, 3060.

31. Humphreys, I. R.; Pei, J.; Baek, M.; Krishnakumar, A.; Anishchenko, I.; Ovchinnikov, Zhang, J.; Ness, T. J. Banjade, S.; Bagde, S. R.; Stancheva, V. G.; Li, X.; Liu, K.; Zheng, Z.; Barrero, D. J.; Roy, U.; Kuper, J.; Fernández, I. S.; Szakal, B.; Branzei, D.; Rizo, J.; Kisker, C.; Greene, E. C.; Biggins, S.; Keeney, S.; Miller, E. A.; Fromme, J. C.; Hendrickson, T. L.; Cong, Q.; Baker, D. Computed Structures of Core Eukaryotic Protein Complexes. *Science*, **2021**, *374*, eabm4805.
32. Zhao, Y.; Rai, J.; Xu, C.; He, H.; Li, H. Artificial Intelligence-Assisted CryoEM Structure of Bfr2-Lcp5 Complex Observed in the Yeast Small Subunit Processome. *Commun. Biol.* **2022**, *5*, 523.
33. Perrakis, A.; Sixma, T. K. AI Revolutions in Biology: The Joys and Perils of AlphaFold. *EMBI Reports*, **2021**, *22*, e54046.
34. Xie, D.; Gulnik, S.; Collins, L.; Gustchina, E.; Suvorov, L.; Erickson, J. W. Dissection of the pH Dependence of Inhibitor Binding Energetics for an Aspartic Protease: Direct Measurement of the Protonation States of the Catalytic Aspartic Acid Residues. *Biochemistry*, **1997**, *36*, 16166-16172.
35. Witt, A. C.; Lakshminarasimhan, M.; Remington, B. C.; Hasim, S.; Pozharski, E.; Wilson, M. A. Cysteine pK_a Depression by a Protonated Glutamic Acid in Human DJ-1. *Biochemistry*, **2008**, *47*, 7430-7330.
36. Loewenthal, R.; Sancho, J.; Fersht, A. R. Histidine-Aromatic Interactions in Barnase. Elevation of Histidine pK_a and Contribution to Protein Stability. *J. Mol. Biol.* **1992**, *224*, 759-770.

37. Hunt, I. Table of pKa and pI Values. University of Calgary, **2021**. https://www.chem.ucalgary.ca/courses/351/Carey5th/Ch27/ch27-1-4-2.html (accessed March 31, 2021).
38. Di Russo, N. V.; Estrin, D. A.; Martí M. A.; Roitberg, A. E. pHDependent Conformational Changes in Proteins and Their Effect on Experimental pKas: The Case of Nitrophorin 4. *PLoS Comput. Biol.* **2012**, *8*, e1002761. doi:10.1371/journal.pcbi.1002761
39. Kasserra, H. P.; Laidler, K. J. pH Effects in Trypsin Catalysis. *Can. J. Chem*. **1969**, *47*, 4021–4029. doi:10.1139/v69-668
40. Antonino, E.; Ascenzi, P. The Mechanism of Trypsin Catalysis at Low pH. Proposal for a Structural Model. *J. Biol. Chem*. **1981**, *256*, 12449–12455.
41. Talley, K.; Alexov, E. On the pH-Optimum of Activity and Stability of Proteins. *Proteins* **2010**, *78*, a–n. doi:10.1002/prot.22786
42. Otto, H.; Marti, T.; Holz, M.; Mogi, T.; Lindau, M.; Khorana, H. G.; Heyn, M. P. Aspartic Acid-96 Is the Internal Proton Donor in the Reprotonation of the Schiff Base of Bacteriorhodopsin. *Proc. Natl. Acad. Sci*. **1989**, *86*, 9228–9232. doi:10.1073/pnas.86.23.9228
43. Woińska, M.; Grabowsky, S.; Dominiak, P. M.; Woźniak, K.; Jayatilaka, D. Hydrogen Atoms Can Be Located Accurately and Precisely by X-ray Crystallography. *Sci. Adv. 2*, **2016**, e1600192. doi:10.1126/sciadv.1600192
44. O'Dell, W. B.; Bodenheimer, A. M.; Meilleur, F. Neutron Protein Crystallography: A Complementary Tool for Locating Hydrogens in Proteins. *Arch. Biochem. Biophys*. **2016**, *602*, 48–60. doi:10.1016/j.abb.2015.11.033

45. Schröder, G. C.; Meilleur, F. Neutron Crystallography Data Collection and Processing for Modelling Hydrogen Atoms in Protein Structures. *JoVE* **2020**, *166*, e61903. doi:10.3791/61903

46. Barrett, P. J.; Chen, J.; Cho, M.-K.; Kim, J.-H.; Lu, Z.; Mathew, S.; Peng, D.; Song, Y.; Van Horn, W. D.; Zhuang, T.; Sönnichsen, F. D.; Sanders, C. R. The Quiet Renaissance of Protein Nuclear Magnetic Resonance. *Biochemistry* **2013**, *52*, 1303–1320. doi:10.1021/bi4000436

47. Bartik, K.; Redfield, C.; Dobson, C. M. Measurement of the Individual pKa Values of Acidic Residues of Hen and turkey Lysozymes by Two-Dimensional ^{1}H NMR. *Biophysical J*. **1994**, *66*, 1180–1184. doi:10.1016/S00063495(94)80900-2

48. Frericks Schmidt, H. L.; Shah, G. J.; Sperling, L. J.; Rienstra, C. M. NMR Determination of Protein pKa Values in the Solid State. *J. Phys. Chem. Lett.* **2010**, *1*, 1623–1628. doi:10.1021/jz1004413

49. Isom, D. G.; Cannon, B. R.; Castañeda, C. A.; Robinson, A.; García-Moreno E., B. High Tolerance for Ionizable Residues in the Hydrophobic interior of Proteins. *Proc. Natl. Acad. Sci.* **2008**, *105*, 17784–17788. doi:10.1073/pnas.0805113105

50. Isom, D. G., Castañeda, C. A., Cannon, B. R., and Garcia-Moreno E., B. Large Shifts in pKa Values of Lysine Residues Buried inside a Protein. *Proc. Natl. Acad. Sci.* **2011**, *108*, 5260–5265. doi:10.1073/pnas.1010750108

51. Bandyopadhyay, D.; Bhatnagar, A.; Jain, S.; Pratyaksh, P. Selective Stabilization of Aspartic Acid Protonation State within a Given Protein Conformation Occurs via Specific “Molecular Association”. *J. Phys. Chem. B* **2020**, *124*, 5350–5361. doi:10.1021/acs.jpcb.0c02629

52. Fitch, C. A.; Karp, D. A.; Lee, K. K.; Stites, W. E.; Lattman, E. E.; García-Moreno, E. B. Experimental pKa Values of Buried Residues: Analysis with Continuum Methods and Role of Water Penetration. *Biophysical J.* **2002**, *82*, 3289–3304. doi:10.1016/s0006-3495(02)75670-1

53. Li, H.; Robertson, A. D.; Jensen, J. H. Very Fast Empirical Prediction and Rationalization of Protein pK_a Values. *Proteins* **2005**, *61*, 704–721. doi:10.1002/prot.20660

54. Spassov, V. Z.; Yan, L. A Fast and Accurate Computational Approach to Protein Ionization. *Protein Sci.* **2008**, *17*, 1955–1970. doi:10.1110/ps.036335.108

55. Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. CHARMM: A Program for Macromolecular Energy, Minimization, and Dynamics Calculations. *J. Comput. Chem*. **1983**, *4*, 187–217. doi:10.1002/jcc.540040211

56. Miao, Z.; Cao, Y. Quantifying Side-Chain Conformational Variations in Protein Structure. *Sci. Rep*. **2016**, *6*, 37024. doi:10.1038/srep37024

57. Ahmed, M. H.; Koparde, V. N.; Safo, M. K.; Neel Scarsdale, J.; Kellogg, G. E. 3D Interaction Homology: The Structurally Known Rotamers of Tyrosine Derive from a Surprisingly Limited Set of Information-Rich Hydropathic Interaction Environments Described by Maps. *Proteins* **2015**, *83*, 1118–1136. doi:10.1002/prot.24813

58. Ahmed, M. H.; Catalano, C.; Portillo, S. C.; Safo, M. K.; Neel Scarsdale, J.; Kellogg, G. E. 3D Interaction Homology: The Hydropathic Interaction Environments of Even Alanine Are Diverse and Provide Novel Structural Insight. *J. Struct. Biol.* **2019**, *207*, 183–198. doi:10.1016/j.jsb.2019.05.007

59. AL Mughram, M. H.; Catalano, C.; Bowry, J. P.; Safo, M. K.; Scarsdale, J. N.; Kellogg, G. E. 3D Interaction Homology: Hydropathic Analyses of the "π-Cation" and "π-π"

Interaction Motifs in Phenylalanine, Tyrosine, and Tryptophan Residues. *J. Chem. Inf. Model*. **2021**, *61*, 2937–2956. doi:10.1021/acs.jcim.1c00235

60. Catalano, C.; AL Mughram, M. H.; Guo, Y.; Kellogg, G. E. 3D Interaction Homology: Hydropathic Interaction Environments of Serine and Cysteine Are Strikingly Different and Their Roles Adapt in Membrane Proteins. *Curr. Res. Struct. Biol.* **2021**, *3*, 239–256. doi:10.1016/j.crstbi.2021.09.002

Chapter 2:

3D Interaction Homology: Computational Titration of Aspartic Acid, Glutamic Acid, and Histidine Can Create pH-Tunable Hydropathic Environment Maps†

Introduction

As mentioned previously, successful molecular modeling studies depend on starting with accurate protein structural data, which can be complicated by hydrogen position uncertainty from crystallographic experiments. Our lab has also previously examined this problem using our inhouse force field HINT (Hydropathic INTeractions)[2-4] that, briefly, exploits experimental libraries of data for atomistic partial $\log P_{o/w}$ values of small molecules and residues to account for enthalpic, entropic, and solvation contributions to free energy and score protein-ligand, protein-protein, protein-nucleotide, etc. interactions. In one study, HINT was used to predict the degree of protonation of ligand-active site interactions of neuraminidase-inhibitor complexes using a method that we termed "computational titration."[5] By scoring all potential models, i.e., where the number of protons attached to ionizable residues and ligand functional groups were exhaustively enumerated, lower energy models were identified. Since proton positions are not unambiguously known from experiment, we term all such models "isocrystallographic" in that all would fit the available electron density envelope.

† This chapter been adapted from Herrington, N. B.; Kellogg, G. E. **2021**[1] Let it be assumed that any indicated supplementary material here is indicated with its name as it appears in Herrington, N.B.; Kellogg, G. E. **2021**.

In another report, HINT modeled the protonation state of a peptide inhibitor–HIV-1 protease complex with pH-dependent interaction scores that paralleled experimental pH-dependent binding data.[6]

Clearly, the presence or absence of protic hydrogens on these residue types within a protein will impact the interactions that these residues make, and in turn the protein's 3D structure. For example, the interaction between two aspartates is radically different if one of the pair is protonated and the proton is oriented to form a hydrogen bond between them. Evaluating and understanding these phenomena is part of our long-term goal of building a new paradigm for protein structure elucidation and prediction.

In the current report, we focus our attention on the hydropathic environments of aspartic acid, glutamic acid, and histidine, three residue types considered to be "ionizable", extracted from the same relatively large dataset of X-ray crystallographic protein structures. Following the same logic used in our previous work, we believe that, not only are each of these residues likely to make their own unique sets of interactions that can be clustered, but their environments also determine each residue's unique ionization state. Thus, using our scoring methods, we have simulated titration of thousands of each of these ionizable residue types to model their protonation in available crystal structures by computationally varying pH. We have generated interaction maps similar to those in our reports on tyrosine,[7] alanine,[8] phenylalanine, tryptophan,[9], serine, and cysteine,[10] but with each possessing an individually optimized protonation state. The role of sidechain buriedness was examined using a calculated solvent-accessible surface area for each of the extracted residues. Further, we show that each residue's backbone

conformation plays a significant role in determining these protonation states. With these, we can directly predict a specific residue's ionization state, explore the effects of varying pH, i.e., tuning, on their hydropathic environments, and collect 3D interaction-similar residue environments by clustering. Moreover, we highlight the most common environments that contribute to one state or another, but more importantly we have developed a basis set of 3D backbone-dependent residue interaction profiles for these three residues that are pieces of the protein structure analysis and prediction puzzle.

Materials and Methods

Dataset

From a collection of 2,703 randomly selected proteins from the RCSB Protein Data Bank, using only structures containing no ligand or cofactor, we extracted all ASP, GLU, and HIS residues from each structure, excluding N- and C-terminal residues. For these structures, we have previously described our selection criteria.[7] Our intention was to abide by random population-based sampling of a variety of primary, secondary, and tertiary structures, thus not excluding proteins with similar or identical sequences. We believe the size of our dataset should exhaust all unique residue environments of HIS, ASP, and GLU. Hydrogen atoms were added to all heavy atoms of all structures based on their hybridization states. Positions of these atoms underwent conjugate gradient minimizations.

Alignment Calculations

We overlayed an 8 by 8 "chessboard" on the standard Ramachandran plot (Figure 1), where each "chess square" has dimensions of 45° by 45° in ϕ (phi)–ψ (psi) space. The

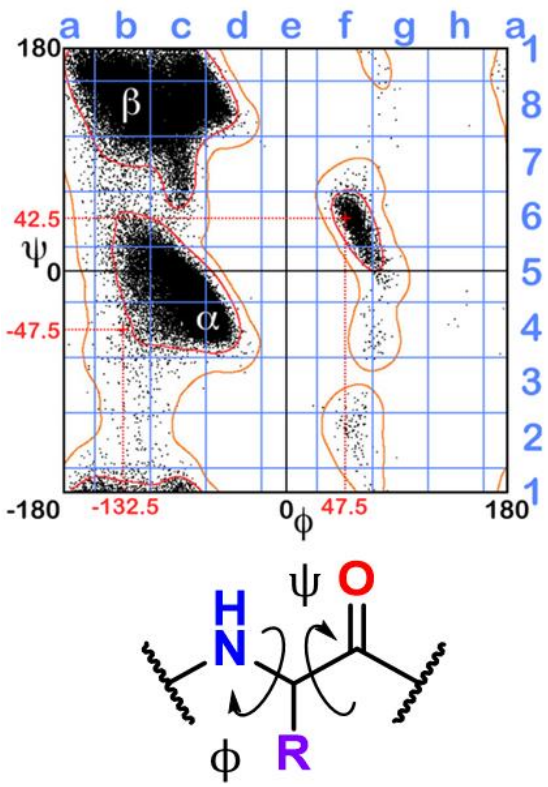


Figure 2.1. Ramachandran plot divided into an 8x8 "chessboard",[27] where individual chess squares have coordinates in letter-number pairs. Residues are binned into chess squares, according to their backbone ϕ (phi) and ψ (psi) dihedral angles.

grid of the board was shifted by −20° and −25° in the ϕ and ψ directions, respectively, to enclose higher-density regions of the plot within single squares. The ϕ, ψ, and χ angles were all calculated for every residue in our dataset, and each residue was binned into their proper chess square based on its respective ϕ and ψ angles. All residues in each chess square were further divided by their χ1 angles into three parse groups: group "0.60", (0° ≤ χ1 < 120°), group "0.180" (120° ≤χ1 < 240°), and group "0.300" (240° ≤χ1 < 360°) (Figure 2.1). In the case of GLU, residues were still further parsed by their χ2 angles, yielding a total of nine parses for this residue. Supplementary Table S1[1] contains all information for each residue of each type in our dataset, including their chess squares, parses, PDB IDs, ϕ, ψ and ω torsion angles and atom numbers for the backbone atoms and CB of each residue.

A single model residue of each type was constructed at the center of each chess square with characteristic ϕ and ψ angles for that centroid. The CA of the peptide backbone was placed at the origin with the CA-CB oriented along the z-axis and the CA-HA bond oriented into the -y, -z quadrant of the yz-plane. All residues of each type were aligned to this model, and rotation and translation matrices were calculated by least-

squares fitting of the residue constituent atoms to the model. This effectively shifted coordinates of every protein structure to align the residue of interest with the centroid within a common frame and ensures that all calculated maps and environments are attributable to a residue's interactions and not misalignments in backbone structure. The average root-mean square distances (RMSDs) for superimpositions of backbone atoms in each chess square are close to 0.15 Å, indicating that errors arising from aligning residue backbones to the centroid model (based on the CA-CB bond) are minimal.

HINT Scoring Function

The HINT forcefield (see Chapter 1)[2-4] was used for all scoring of interactions between protein atoms. HINT relies on atom-focused parameters, namely the hydrophobic atom constant (a_1) and a value for solvent-accessible surface area (SASA, S_i) for atom *i*. Generally speaking, $a_i > 0$ for hydrophobic atoms and $a_i < 0$ for polar atoms.

S_i is greater for more solvent-exposed external atoms. The interaction score between atoms i and j is calculated by:

$$b_{ij} = a_i S_i a_j S_j T_{ij} \, e^{-r} + L_{ij},$$

where r is the distance in angstroms between atoms i and j. T_{ij} is equivalent to −1, 0, or 1 to account for acidic, basic, etc. character of atoms involved and assign the proper sign to the interaction score. Finally, L_{ij} implements the Lennard-Jones potential function.[2] $b_{ij} > 0$ for favorable interactions, such as Lewis acid-base and hydrophobic-hydrophobic interactions, while $b_{ij} < 0$ for unfavorable interactions, including hydrophobic-polar or Lewis base-base interactions.

Computational Titration of Ionizable Residues

To determine the optimal ionization state of each studied residue, we adapted an algorithm that we reported previously for improving protein-ligand models for scoring.[2,4,11] Our algorithm scores all possible ionization states of a model residue with other residues in its environment. Here, we optimized the ionization states of residues by first calculating the normal (environment-free) cost for ionizations of these residues using published data (ASP, pK_a = 3.65; GLU, pK_a = 4.25; HIS, pK_{a1} = 6.00, pK_{a2} = 14.44)[12] and applying the Henderson-Hasselbalch equation. For ASP, at pH 7, log [(CO_2^-]/(CO_2H)] = 3.35, which is an equilibrium constant that can be converted to a ΔG of 4.57 kcal mol^{-1}. Using the previously reported relation that −1 kcal mol^{-1} ≈ 500 HINT score units, the energy cost in HINT score units for protonating aspartate at pH 7, in the absence of local pH effects is 2,295. Table 2.1 summarizes these energy costs.

Table 2.1. Energy costs in HINT scores for computational titration of aspartic acid, glutamic acid and histidine at various pH values.

	pK_a	pH 4	pH 5	pH 6	pH 7	pH 8	pH 9	pH 10
Aspartic Acid	3.65	240	925	1610	2295	2980	3665	4350
Glutamic Acid	4.25	-171	514	1199	1884	2569	3254	3939
Histidine K_{a1}	6.00	-1370	-685	0	685	1370	2055	2740
Histidine K_{a2}	14.44	7151	6466	5781	5096	4411	3726	3041

The second term, calculated for each residue in varying protonation states, also as a HINT score, measures the effects of the local environment around the residue. This assessment of the environment scores the interactions of the residue in question with those nearby, in each accessible protonation state. These scores are summed together with the appropriate values in Table 2.1 to determine the best scoring, and therefore most likely, protonation state of the residue. For ASP and GLU, we examined the ionized (carboxylate, CO_2^-) and neutral states with protonation at each oxygen atom (OD1/OE1

and OD2/OE2). For the latter, the -C-C-O-H dihedral angles were exhaustively optimized for ideal hydrogen bonding to surrounding residues. For HIS, four potential ionization states exist: 1) protonation at both ND1 and NE2 (HIS^+), 2) protonation at only ND1 (HIS-δ), 3) protonation at only NE2 (HIS-ε) and 4) deprotonated (HIS^-), the last of which is reported to be exceedingly rare. Since the entire imidazole ring of HIS can be flipped, the potential cases for this residue are doubled to eight (vide infra). If the HINT score was 50 or more (~0.1 kcal mol^{-1}) than the starting case, the residue's molecular model was replaced with the (protonated or deprotonated) trial model for that case. All further calculations at that pH were performed with the resulting optimized residue structure and coordinates.

pK_a Calculations

We identified 94 residues with experimental pK_a values in the PKAD database[13] that were also present in our dataset and compared our predicted pK_a values for those to their experimental values. Using the technique described above, we calculated individual pK_a values for these residues and compared them with those in the PKAD database. Calculation of a residue's protonation state was performed within a range from 1 to 14 in increments of a quarter of a pH unit. We treated the two points representing the protonation transition state as part of a linear regression and solved for the "equivalence point" between them.

HINT Basis Interaction Maps

Each residue with its CA-CB bond along the z-axis, was placed within a three-dimensional box large enough to accommodate the structure of a residue, plus an

additional 5Å on each dimension. These boxes, based on residue type, are as follows: ASP, –8.5 Å ≤ x ≤ 8.5 Å; –8.5 Å ≤ y ≤ 8.5 Å; –7.5 Å ≤ z ≤ 9.5 Å, (42,875 points, 4,913 Å^3); GLU, –8.5 Å ≤ x ≤ 8.5 Å; –8.5 Å ≤ y ≤ 8.5 Å; –7.5 Å ≤ z ≤ 10.5 Å, (45,325 points, 5,202 Å^3); and HIS, –10.0 Å ≤ x ≤ 10.0 Å; –10.0 Å ≤ y ≤ 10.0 Å; –7.5 Å ≤ z ≤ 9.5 Å, (58,835 points, 6,800 Å^3); all with a point spacing of 0.5 Å. As described previously,[27] HINT was used to calculate an interaction grid representing the 3D interaction space surrounding a residue of interest. In short, these maps interpret sums of pairwise HINT scores[2-4] into 3D map objects indicating position, intensity, and type of interaction between atoms of the residue and those close in proximity. Each grid point for a map was calculated, according to:

$$\rho_{xyz} = \Sigma\, b_{ij} \exp\{ -[(x - x_{ij})^2 + (y - y_{ij})^2 + (z - z_{ij})^2] / \sigma\},$$

where ρ_{xyz} is the map interaction score at coordinates (x, y, z), x_{ij}, y_{ij} and z_{ij} are coordinates of the midpoint of the vector between atoms *i* and *j*, and σ is the width of the Gaussian map peak, 0.5 for our purposes.[7] Map data were calculated for sidechain atoms of all ASP, GLU, and HIS residues with individual maps for the four interaction classes: favorable/unfavorable polar and favorable/unfavorable hydrophobic.

Calculation of Map-Map Correlation Metrics

Comparison of two maps, *m* and *n*, are based on:

$$\text{If } |Gt| / F > 1.0, \quad A_t = (G_t / |G_t|) \log_{10}(|G_t| / F); \text{ else, } A_t = 0,$$

where each raw map data point (G_t, for point at index t) is transformed to $\log_{10}$ space and normalized with a predefined floor value, F = 1.0. Similarity between maps *m* and *n*, defined as *D*(m,n) is calculated based on previous methods:[7]

$$D(\mathrm{m},\mathrm{n}) = \Sigma \{1 - (|A_t(\mathrm{m}) - A_t(\mathrm{n})|)^2 / [(|A_t(\mathrm{m})| + |A_t(\mathrm{n})|) \cdot (|A(\mathrm{m})|_{\max} + |A_t(\mathrm{n})|_{\max})]\}.$$

In this metric, $A_t(\mathrm{m})$ and $A_t(\mathrm{n})$ are map values for the same grid points in maps m and n, respectively, and $|\mathrm{A}|_{\max}$ is the absolute max value of the grid points in *m* and *n*. Our map boxes are designed to accommodate all possible residue environments and usually contain a majority (>60%) of zero-valued points. To mitigate the issue that all map pairs would appear similar, only points where $|A_t(\mathrm{m})| \geq 8\,|A(\mathrm{m})_{\mathrm{stddev}}|$ or $|A_t(\mathrm{n})| \geq 8\,|A(\mathrm{n})_{\mathrm{stddev}}|$ (A_{stddev} is the standard deviation of the average value of all points in the map) in calculating $D(\mathrm{m},\mathrm{n})$[7] were considered.

$D(\mathrm{m},\mathrm{n})$ should normally range from 0 to 1, where 1 indicates identical maps; realistically, $D(\mathrm{m},\mathrm{n}) = 0$ cannot exist, as it would signify completely overlapping maps with opposite signs. Neither will $D(\mathrm{m},\mathrm{n}) = 0.5$ exist, as it would require completely non-overlapping maps. Typically, the minimum *D* thus falls between 0.6 and 0.7. To calculate the overall similarity (D_{all}) between two like residue maps m and n, one composite metric was calculated from four metrics containing data for the map quartet described above [hydro (+), hydro (−), polar (+), and polar (−), which are favorable and unfavorable hydrophobic (e.g. hydrophobic-polar) contributions, and favorable and unfavorable polar contributions to each map, respectively]. Here, $D(\mathrm{m},\mathrm{n})_{\mathrm{all}} = \{4[D(\mathrm{m},\mathrm{n})_{\mathrm{hydro(+)}}] + 2[D(\mathrm{m},\mathrm{n})_{\mathrm{hydro(-)}}] + [D(\mathrm{m},\mathrm{n})_{\mathrm{polar(+)}}] + [D(\mathrm{m},\mathrm{n})_{\mathrm{polar(-)}}]\} / 8$.

The favorable and unfavorable hydrophobic interactions were scaled by 4 and 2, respectively; these two terms are more subtle, diverse, and potentially information-rich, than those driven by electrostatic, particularly ionic, interactions.

Also, to reduce the computational burden, we applied a first-pass similarity filter[27] to our matrix calculations to remove certain residues from further consideration because

many maps are highly similar as they share highly similar environments, and thus can be removed to avoid redundancy. This significantly scales down our pool of calculations, which is significant as several steps scale more or less as n^2.

As described previously,[7] all above calculations were performed with in-house-written programs that exploit the inherent parallelism of our methods with GPUs, specifically used to calculate maps and similarity matrices.

Clustering and Validation

We utilized the freely available R programming language and environment[34] to perform our clustering analysis on the pairwise map similarity matrices calculated above. We determined[27] that for our purposes, out of a number of different clustering methods, the k-means method was most reliable. Through the experience of our previous reports[7,8] and preliminary studies here, we opted to set a uniform maximum number of clusters of 12 for each chess square-parse combination. This allows for significant map diversity and facilitates inter-chess square/ inter-residue comparisons. Most chess squares/parses, however, had fewer than 12 clusters in their optimal solutions. Additionally, k-means clustering will not form singleton clusters, i.e., with a single member. However, while this is fairly rare (~5%), these maps could be interesting, so our protocols are designed to optionally recover them by reconstructing the cluster solutions with the missing singletons. Any chess square-parse with four or fewer maps was not clustered, but, instead, averaged to create what is, effectively, a 1-cluster case.

Average Map, RMSD, and Solvent-Accessible Surface Area Calculations

Careful consideration must be given to calculation of average maps. First, to avoid what we have described as "brown mapping,"[7] only maps sharing high similarity should

be combined. Second, the average maps are calculated by Gaussian weighting (w) the contribution of each map with respect to its Euclidean distance from the cluster centroid, given by:

$$w = \exp\left[-\left(d^2/\sigma^2\right)\right],$$

where d is the map's distance from the centroid and $\sigma = d_{max}/8$, which is the average of all maximum distances across all clusters in the chess square. This weighting ensures that maps closer to the centroid contribute more significantly to the average map of the cluster, whereas taking a flat average of all map data would overweight the importance of maps further from the centroid. While a formal definition exists for "exemplar" in affinity propagation clustering, for our purposes, it represents the residue datum closest to the centroid of each cluster output by the k-means algorithm.

RMSDs (root-mean square distances) for each residue type were calculated by weighted averaging, as above, all atomic positions from all residues in a cluster to construct one average residue structure. For each non-hydrogen atom, an RMSD was calculated from the average structure, and then all atomic values were averaged to obtain the reported RMSD for the cluster.

We calculated SASAs for all residue sidechains using the GETAREA algorithm[35] and its default settings. The protein coordinates in PDB files were submitted as input. Also from GETAREA's "In/Out" parameter, we created a new metric "$f_{outside}$" to represent the buriedness of the set of residues in a cluster, parse, chess square, etc. by recasting "In" as 0.0, "Out" as 1.0 and "indeterminant" as 0.5, and averaging the set.

Results and Discussion

Dataset: Binning and Parsing Residues

From the dataset of 2,703 protein structures described in Methods, we extracted 42,713 ASPs, 49,306 GLUs, and 15,276 HISs, all of which were non-terminal residues. An 8 by 8 chessboard was overlaid on a standard Ramachandran plot,[16] such that each grid square has dimensions of 45° by 45° in ϕ–ψ space and the extents of the board are shifted slightly to contain regions of high residue population density in single squares (Figure 2.1), named as a1 through h8. We binned residues into each square by their backbone ϕ and ψ angles and further parsed them by their χ1 angles into three groups corresponding to those normally observed in rotamer libraries:[17] a group averaging ~60°, a group averaging ~180°, and a group averaging ~300° from here on referred to as the "0.60", "0.180", and "0.300" parses. In the case of GLU, residues were still further parsed by their χ2 angles, yielding a total of nine parses for this residue: "0.60.60", "0.60.180", "0.60.300", "0.180.60", "0.180.180", "0.180.300", "0.300.60", "0.300.180" and "0.300.300" (Figure 2.2). We showed previously[7] that map-based clustering was able to easily identify this

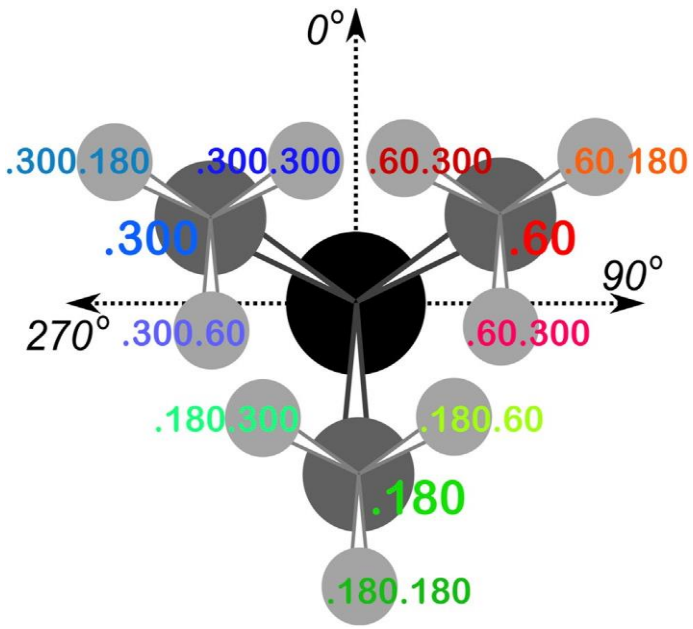


Figure 2.2. The χ_1 and χ_2 rotamer parses. CB (black) has three χ_1 rotamers (dark gray, CG): 0.60, 0.180, 0.300. Each of those, for GLU, has three χ_2 rotamers (light gray, CD), as shown.

(χ1, χ2) low level of detail, except for surface-exposed residues that show few interactions with anything apart from solvent. However, even a few such failures were problematical in calculating average maps and residue coordinates. Furthermore, parsing of the chess square members into χ bins increased computational efficiency. (Many calculations scale as n^2: $3 \times (n/3)^2 < n^2$). The additional χ2 parse for GLU further reduced the computations and made the ASP and GLU data more comparable, i.e., the (unparsed) remainder of their sidechains is the same –C–COOH fragment.

Throughout this work, chess square names will be given in bold italics, e.g., a1, b4, etc. The χ1 parses for ASP and HIS will be denoted by the suffixes 0.60, 0.180 and 0.300 and the χ1/χ2 parses for GLU will be denoted by the suffixes 0.60.60, 0.60.180, 0.60.300, etc.

The occupancies of the chess square/parses range from 0 to 6,215 (d4.300) for aspartate, to 4,563 (d4.300.180) for glutamate, and to 1,504 (d4.180) for histidine. For aspartate, 44 (of 64) chess squares contain 10 or more residues, and 159 chess squares/ parses (of 192) are occupied at all. These metrics are 40/64 and 356/576 for glutamate and 32/64 and 120/192 for histidine. Table 2.2 provides occupancies in the Ramachandran chessboards for these three residues. To simplify nomenclature in this article, we are using a numerical scheme wherein the sequential number of that residue in its chess square/parse is its name. Thus, histidine 100 in chess square a1.60 is the 100th histidine contained within that chess square/parse combination, as tabulated in Supplementary Table 2,[1] wherein the specific actual PDB ID, chain, residue name, etc. for

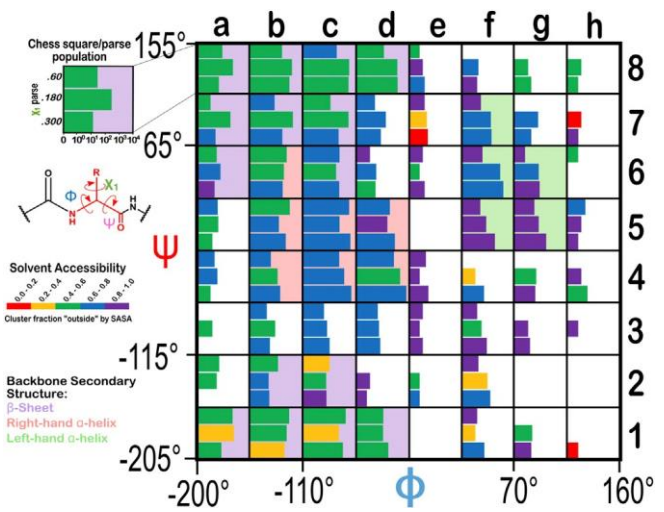


Figure 2.3. Ramachandran chessboard displaying the chess square/parse population for aspartic acid. The Ramachandran ϕ vs. ψ plot is rendered into 64 45° by 45° (π/4 by π/4) chess squares. The (χ_1) parse populations for ASP are represented in $\log_{10}$ scale with the colored bares. Their colors reflect the average weighted fraction outside or solvent-exposed, i.e. "f_{outside}," a measure of solvent accessibility (see text for definition). The ϕ vs. ψ regions associated with β-pleat, α-helix, and left-hand α-helix secondary structure motifs are shaded in light purple, light orange, and light green chess squares, respectively.

each datum in this study can be found. Clusters (vide infra) will be named for the residue closest to its centroid or exemplar and will be given in bold numerals.

The Ramachandran plot generally contains four regions associated with specific secondary structure motifs. Figure 2.3 shows an example of a fully-binned Ramachandran plot, in this case, for aspartic acid. According to our schema, fifteen chess squares (a1, a6, a7, a8, b1, b2, b7, b8, c1, c2, c6, c7, c8, d1 and d8) correspond to the β-pleat motif, seven chess squares (b4, b5, b6, c4, c5, d4 and d5) correspond to the right-hand α-helix motif and five chess squares (f5, f6, f7, g5 and g6) correspond to the left-hand α-helix motif. The remaining chess squares, some of which may contain mixtures of secondary structural motifs, account for the remaining residues.

Table 2.2. Number of residues in each chess square and parse for ASP, GLU and HIS.

	Number of aspartates in parses			**60 / 180 / 300**				
	a	***b***	***c***	***d***	***e***	***f***	***g***	***h***
1	261/342/22	556/323/200	1162/260/650	39/37/120	0/0/0	3/2/15	0/5/4	0/0/1
2	14/8/0	47/6/7	33/16/13	0/2/1	0/1/1	4/29/56	0/0/0	0/0/0
3	0/3/0	4/25/8	10/20/34	11/15/19	2/1/2	3/8/25	0/2/3	0/1/0
4	5/9/2	10/44/76	279/833/4142	395/1578/6215	4/1/7	0/2/14	0/12/10	0/2/8
5	10/11/3	647/60/263	2539/309/3409	169/92/477	0/0/0	6/22/160	8/21/111	5/2/1
6	8/18/5	326/181/131	313/142/274	2/5/7	2/1/3	10/492/1035	1/21/25	1/0/0
7	2/163/6	23/1279/126	43/2250/254	6/67/23	3/5/6	7/65/74	0/18/4	0/2/1
8	27/292/13	118/1035/615	168/2483/2284	44/854/919	1/2/3	0/4/3	0/2/4	0/2/1
	Number of glutamates in parses			**60.60 / 60.180 / 60.300 180.60 / 180.180 / 180.300 300.60 / 300.180 / 300.300**				
	a	***b***	***c***	***d***	***e***	***f***	***g***	***h***
1	4/218/15 16/24/2 2/22/4	20/420/33 30/65/6 28/490/233	7/95/35 35/40/2 66/538/279	1/18/12 13/16/3 22/50/24	0/0/0 0/0/0 0/0/0	0/0/0 0/0/0 3/3/0	0/0/0 0/0/0 0/0/0	0/0/0 1/0/0 0/0/0
2	0/0/0 0/3/0 0/1/0	3/8/1 1/0/0 3/10/6	0/2/0 3/0/1 4/9/6	0/0/1 0/1/0 1/1/1	0/0/0 0/0/0 0/0/0	1/1/0 3/1/0 2/25/17	0/0/0 2/0/0 0/0/0	0/0/0 1/0/0 0/0/0
3	0/0/0 0/1/0 0/0/0	0/1/0 1/10/1 1/8/2	0/1/0 12/14/2 10/18/5	2/2/2 18/19/2 9/23/14	0/1/0 5/4/0 2/1/3	0/0/0 1/3/0 3/7/1	0/0/0 2/0/0 1/4/2	0/0/0 1/0/0 0/0/0
4	1/4/1 5/8/3 0/3/2	1/10/1 22/43/9 7/95/58	25/299/133 1135/1983/224 698/3463/1799	32/374/212 1393/4345/325 1139/4563/1868	0/3/1 2/2/1 2/2/1	0/1/0 1/1/0 0/3/3	0/0/0 0/8/0 0/5/2	0/1/0 4/2/1 0/0/0
5	1/4/1 2/0/0 1/4/0	1/62/10 14/30/3 23/261/87	41/538/438 174/252/25 465/2053/1089	11/219/153 54/122/9 152/311/122	0/0/0 0/0/0 0/0/0	0/3/2 1/4/0 11/56/19	0/0/0 2/2/0 6/46/21	0/0/0 0/0/0 0/0/0
6	0/0/1 2/6/0 0/4/0	1/41/3 11/12/3 6/111/34	5/42/15 20/19/8 30/119/68	0/1/1 0/1/1 1/1/1	0/2/0 1/0/0 0/2/1	2/5/3 27/23/0 17/378/145	0/0/0 3/1/0 2/7/2	0/0/0 0/0/0 0/0/0
7	0/2/0 6/15/3 0/3/1	2/17/1 64/201/17 18/223/78	2/15/7 69/208/28 38/262/109	0/0/2 11/16/3 2/16/8	0/4/1 1/0/0 2/3/0	0/6/1 15/2/1 4/33/11	0/1/0 2/3/0 0/0/3	0/0/0 1/4/1 0/0/1
8	10/84/4 61/234/10 0/23/10	15/442/25 251/1511/55 56/1852/419	19/213/56 404/1854/84 181/1733/779	14/91/23 193/645/43 87/376/217	0/0/0 1/3/2 2/0/1	0/0/0 1/6/0 1/1/0	0/1/0 0/3/0 3/1/2	0/1/0 0/0/0 0/0/0
	Number of histidines in parses			**60 / 180 / 300**				
	a	***b***	***c***	***d***	***e***	***f***	***g***	***h***
1	192/22/3	316/27/416	119/35/226	14/10/15	0/0/0	1/2/1	0/1/0	0/0/1
2	1/0/0	4/2/8	0/0/5	0/0/0	0/0/0	0/0/12	0/0/0	0/0/0
3	0/0/0	6/11/13	1/7/7	0/13/3	0/1/0	0/2/4	0/0/0	0/0/0
4	1/5/1	4/30/80	94/739/947	125/1504/905	0/1/1	0/3/4	0/7/5	0/0/0
5	2/0/7	45/10/431	395/69/1249	98/32/86	0/0/0	0/1/55	0/1/42	0/0/0
6	1/2/2	25/12/282	9/17/206	0/1/0	0/0/0	20/43/503	0/0/24	0/0/0
7	1/14/0	12/176/251	3/231/259	4/3/3	2/1/1	5/18/17	0/1/5	0/0/1
8	54/229/7	172/759/1009	39/748/1025	34/375/171	0/1/2	0/0/3	0/1/9	2/1/0

Calculations in this study were performed for all Ramachandran chess squares, but, for brevity's sake, we focus our discussion on a particular four, designed to sample the three major regions of the standard Ramachandran plot: b1, c5, d5 and f6. The c5, d5 pair allows us to compare independently-calculated map and environment data between chess squares within the same right-hand α-helix structural motif region.

Ionization State Optimization

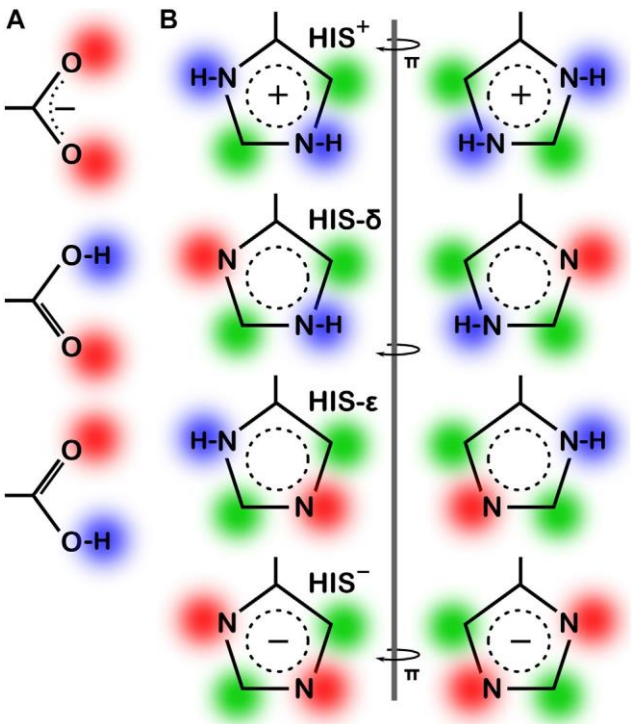


Figure 2.4. Various possibilities for ASP, GLU, and HIS ionization/rotameric states. A) ASP, GLU, and B) HIS sidechain functional groups. Red = Lewis acid, blue = Lewis base, green = hydrophobic. Note that "ring flips" of HIS present distinct patters for interaction.

While our primary goal for this study is to evaluate the hydropathic environments of the ASP, GLU and HIS residue types, a key requirement was to use molecular models that are in appropriate ionization states. We were also interested in examining the effects of these ionization states on the residue environments. Also, such structures (and 3D maps) should have rational and tunable pH dependencies to enable prediction of structure, properties, and function.

As the local environment heavily influences protonation states of ionizable residues, we updated the computational titration algorithm that we reported earlier[23,25] to optimize the ionization state (and concomitantly the–C–O–H dihedral angle) of all

residues in this study. Briefly (Methods), we calculated the HINT score between each residue and its local environment in each of its possible ionization/rotameric states (3 for ASP and GLU, 8 for HIS, Figure 2.4). These scores were modified by pK_a- and pH-dependent factors derived from the Henderson-Hasselbalch equation. It is important to emphasize that all these calculations were performed without changing the atomic positions of the non-hydrogen atoms—except for the π rotation about χ2 shown on the right side of Figure 2.4B. In other words, all models generated and scored are isocrystallographic. The highest-scoring model of the set generated for each residue was selected for moving forward in the study. We note an advantage here: since the positions of the heavy atoms are fixed based on their X-ray structures, calculations will likely identify the protonation model most favorable for that conformation.

Aspartic Acid

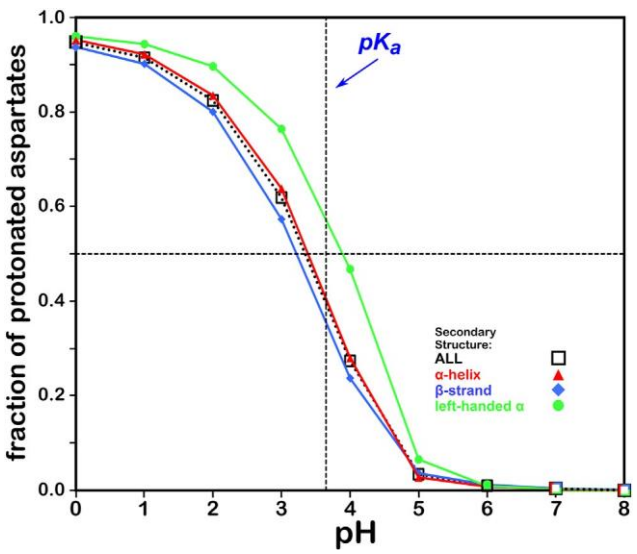


Figure 2.5. Titration curves of ASP residues by secondary structure. The native pK_a for aspartic acid is indicated.

We calculated the optimal structure for each studied aspartic acid at a range of pHs. For this residue, where the pK_a is 3.65, we determined the fraction of the nearly 43,000 residues protonated at pHs from 0 through 8. The result, which is reminiscent of a titration curve, is shown in Figure 2.5. Our calculations yielded the total fraction of aspartic acids expected to be protonated at pHs 0 through 8 in increments of 1 with an overall titration curve centered close to the nominal ASP pK_a

and differing, overall, by ~0.31 pH units. Our calculations suggest that residue backbone structure has an impact on levels of protonation. Our data (vide infra) also suggest that differences in secondary structure have an effect on solvent accessibility: these two phenomena are intimately linked, and in fact difficult to separate. pK_a shifts associated with differences in solvent-accessible surface area are known, as less solvent exposure may increase the pK_as of acidic residues.[18] Highly solvent-exposed residues are, in practice, in vacuo in many protein structure models so that there are no inter-residue interactions to account for. The pH in our calculations at which the aspartic acids are 50% ionized (which we are calling pH_{50}) is 3.345. While this is an arbitrary value, we will use pH_{50}s as set points for map calculations (see below).

Glutamic Acid

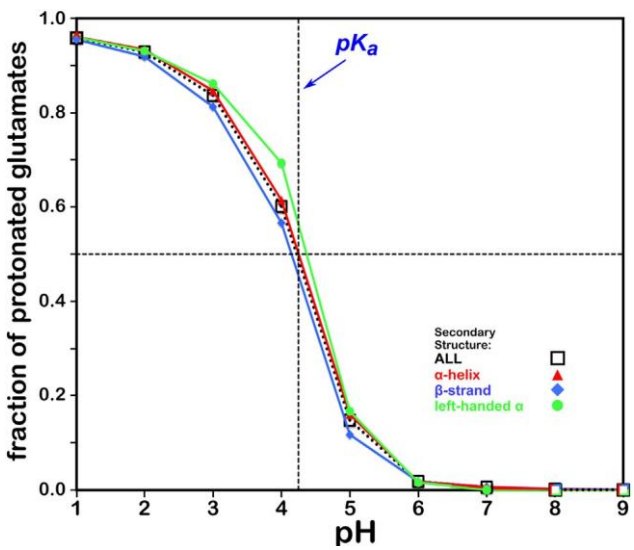


Figure 2.6. Titration curves of GLU residues by secondary structure. The native pK_a for aspartic acid is indicated.

The titration curves for the over 49,000 GLU residues in our study are shown in Figure 2.6. These look very similar to those of ASP and, in the same way, center very closely to its native experimental pK_a. In fact, the average calculated GLU pK_a deviated from the experimentally-determined pK_a for the GLU model peptide by only ~0.03 pH units. There is also seemingly less secondary structure dependence for these results, which is likely due to differences in solvent accessibility between ASP and GLU sidechains. pH_{50} for our glutamic acid data is 4.224.

Histidine

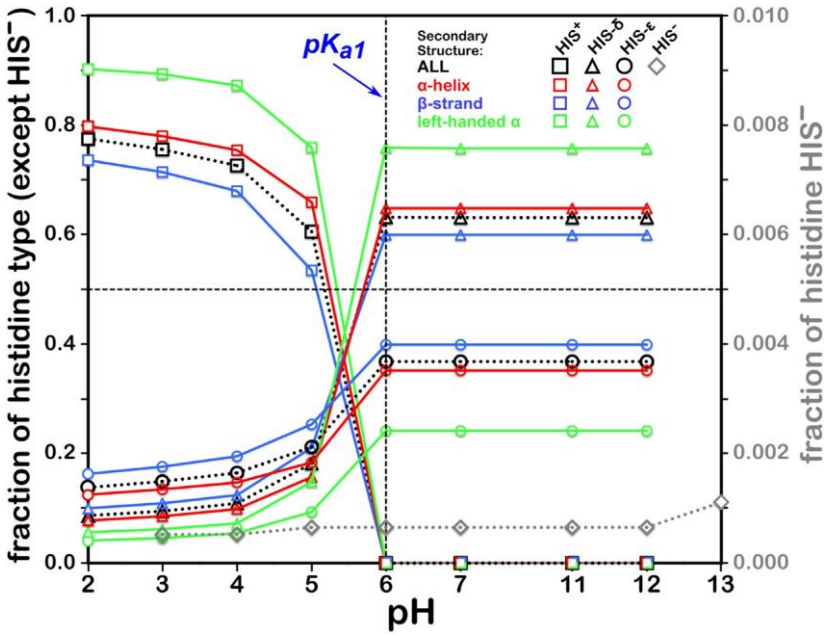


Figure 2.7. Titration curves of HIS residues by secondary structure. The native pK_a1 for histidine is indicated. Full deprotonation of HIS to HIS^- is shown with data colored in gray and right-hand y-axis.

This residue type potentially has three different protonation states, resulting in four unique protonation patterns (Figure 2.4), compared to ASP's and GLU's two, and thus tells a more complicated story (Figure 2.7). In addition to the expected HIS to HIS^+ protonation, HIS can be deprotonated to HIS^-[19] in exceedingly rare cases, such as Cu, Zn superoxide dismutase. We simulated the titration of more than 15,000 HIS residues in our dataset together and separately by their secondary structure. According to our calculations, in the neutral state, a greater fraction of HIS residues were protonated at the ε-nitrogen in all secondary structures. However, factors contributing to protonation of HIS are much more complicated, including solvent accessibility and conformational changes, discussed later. The deviation of our calculated pH_{50} of 5.174 from the nominal HIS pK_{a1} of 6.00 is greater for HIS than those of ASP and GLU, here ~0.83 pH units. Also interesting is that apparently only around 80% of HIS residues can even be protonated to HIS^+, likely due to steric constraints disallowing that configuration, but for HIS in left-

hand α-helix conformations, 90% can be protonated, presumably due to less structural constraint imposed by that backbone motif.

Summary of pH Optimization Results

Although this was a secondary goal, our predictions for residue pK_as are reasonable enough (Supplementary Table 3)[1] that the molecular models upon which our 3D maps are constructed are likely to be correct, as least as snapshots of them in the dynamic biological solution. Our algorithm tends to simulate ionization for highly solvent-exposed residues in protonated forms (charge neutral for ASP and GLU and cationic for HIS). As noted above, there are no interacting residues and (usually) few or no explicit water molecules in the protein models for such residues to aid in the estimation, and the few interactions that are found prefer uncharged species. Our simulation of "bulk" solvent is only through the pressure applied by the external pH term in the Henderson-Hasselbalch relation. For high-level pK_a estimations, clearly more rigorous consideration of solvent molecules and, as Friedman showed,[20] ions, may provide more accurate predictions of ionization states. However, on the $\sim 10^5$ case scale of this study, we used our more practical and accessible approach.

Interestingly, the easier to experimentally determine pK_as of surface residues[21] contrasts with the easier to calculate pK_as of more buried residues, and there is not really a lot of experimental data available. The ionization state-optimized molecular models, which are more important for our purposes, are likely to be quite reasonable except in edge cases. The computationally more problematical highly solvent-exposed residues are fully immersed in water and are thus less participatory in protein structure. We will show below that the edge cases, themselves, are also not a significant issue because it is

interactions that are assayed by the maps, and an ASP, GLU or HIS can be a donor and/or an acceptor.

Calculation of Hydropathic Environment Maps

Based on methods in our previous reports[7-9] we evaluated interatomic interactions using the HINT force field and score model,[2-4] which uses two atom-centered parameters a_i and S_i, the partial log $P_{o/w}$ (for 1-octanol and water solute transfer) and a term related to solvent accessible surface area, respectively, for atom *i* to score atom-atom interactions (see Materials and Methods). We have reported previously on HINT's ability to estimate changes in free energy for ligand-protein, protein-protein and other complexes in various systems,[22-25], such that ~500 HINT score units correlate well with a $\Delta\Delta G = -1$ kcal mol^{-1}.

As stated above, one of our primary hypotheses is that there is a limited set of unique 3D hydropathic interaction environments that satisfy the "valence" of a residue. These valences are based on interaction types, strengths, and geometry. For example, as we showed in previous work[7] the phenol hydroxyl of tyrosine can make favorable polar interactions with an appropriately positioned hydrogen bond donor and/or acceptor, and it can take the form of a backbone amide, another polar sidechain, or a water molecule. In contrast, our alanine maps showed fewer unique interactions, with its methyl sidechain and no rotamers, but about four to six specific patterns appeared to be conserved.[8] Consistent in both of these studies is that we only need to be focused on the interactions that a residue makes with its environment by class, not by the specific donor-acceptor pair or residue type identities. In other words, the type of interaction, its strength and location are more significant than its participants.

Maps were constructed within rectangular boxes tailored to be large enough to contain each of our three studied residue types with its interacting atoms (Materials and Methods). These maps are calculated to quantify the strength of the variety of interactions each residue in our dataset makes with the other atoms in its environment. Our maps categorize interactions in "quartets" of four separate types: favorable polar, unfavorable polar, favorable hydrophobic and unfavorable hydrophobic. Our previous work on tyrosine[7] and alanine[8] examined the hydropathic environments as stand-ins for structure. Here, we exploit these maps that encode extensive information concerning the structural roles of the carboxylates and sidechains of aspartate and glutamate and the dual proton acceptor-donor nature of histidine's imidazole. Our map data further use this information to account for the environments that potentially stabilize any of these residue's ionization states, particularly in response to changes in pH.

Evaluating the Fundamental Patterns in the Maps

To extract the information encoded in the 3D hydropathic interaction maps, we first developed a map-map similarity metric[7] to score two maps m and n (section Materials and Methods). In brief, the overall similarity (D_{all}) between two like residue maps *m* and *n*, is comprised of a single scalar metric derived by the linear combination of four terms, one for each member of the map quartet contributions to each map, respectively. These scalars were loaded in square matrices, for each chess square and parse, for statistical analysis. Next, we clustered these matrices with k-means clustering within the R programming environment. As described in Materials and Methods, we set a maximum number of 12 clusters per chess square-parse combination; this was sufficient for capturing the diversity of residue environments while balancing

computational efficiency. Table 2.3 sets out the number of clusters found on a chess square-parse basis for the three residue types in this study.

Hydropathic Interaction Maps

The objective of examining maps is to view 3D representations of the positions and magnitudes of the constellation of interactions made by residues. We expected that secondary structural differences affect the interactions a residue makes with its environment, which we enforced with the chessboard schema. Additionally, the parse inside each chess square may impact these interactions. For these reasons, we focused the analysis presented here on four particular chess squares, b1, c5, d5 and f6, to survey the environments from each of the three secondary structural regions of the Ramachandran plot, as in previous reports.[8,9] We performed complete studies for all three residues at pHs 3, 5, 7, and 9 and at the pH for each residue at which half of all of that type of residue were protonated, which we named pH_{50} above. However, we only constructed visual map contours displays at each residue's pH_{50}, as we believed this pH would be best representative of the diversity of maps in protonated and deprotonated cases.

Aspartic Acid

Aspartic acid, by nature, is an extremely polar residue, owing to its carboxy acid sidechain. For this reason, we expected to see two things: 1) a plethora of maps indicating strong favorable and unfavorable polar interactions localized around the carboxylate end of the sidechain and 2) many clusters of maps with high solvent-accessible surface areas,

Table 2.3. Number of clusters in each chess square and parse for ASP, GLU and HIS.

	Number of aspartate clusters in parses			**60 / 180 / 300**				
	a	***b***	***c***	***D***	***e***	***f***	***g***	***h***
1	12/12/5	12/12/12	12/12/12	9/7/10	0/0/0	1/1/3	0/2/1	0/0/1
2	4/2/0	8/2/2	6/4/3	0/1/1	0/1/1	1/5/11	0/0/0	0/0/0
3	0/1/0	1/5/3	3/5/6	3/5/5	1/1/1	1/2/6	0/1/1	0/1/0
4	1/3/1	3/8/11	12/12/12	12/12/12	1/1/2	0/1/4	0/3/3	0/1/1
5	3/3/1	12/7/12	12/12/12	11/10/12	0/0/0	2/6/12	3/5/12	2/1/1
6	3/5/2	12/12/8	12/11/12	1/2/2	1/1/1	3/12/12	1/5/5	1/0/0
7	1/11/2	5/12/12	7/12/12	2/9/5	1/1/1	2/10/9	0/4/1	0/1/1
8	5/12/4	11/12/12	12/12/12	7/12/12	1/1/1	0/1/1	0/1/1	0/1/1
	Number of glutamate clusters in parses			**60.60 / 60.180 / 60.300** **180.60 / 180.180 / 180.300** **300.60 / 300.180 / 300.300**				
	a	***b***	***c***	***D***	***e***	***f***	***g***	***h***
1	1/12/4 4/4/1 1/5/1	4/12/8 6/7/2 4/12/12	3/12/6 5/6/1 11/12/12	1/4/3 4/5/1 4/9/5	0/0/0 0/0/0 0/0/0	0/0/0 0/0/0 1/1/0	0/0/0 0/0/0 0/0/0	0/0/0 1/0/0 0/0/0
2	0/0/0 0/1/0 0/1/0	1/3/1 1/0/0 1/3/2	0/1/0 1/0/1 1/3/2	0/0/1 0/1/0 1/1/1	0/0/0 0/0/0 0/0/0	1/1/0 1/1/0 1/6/5	0/0/0 1/0/0 0/0/0	0/0/0 1/0/0 0/0/0
3	0/0/0 0/1/0 0/0/0	0/1/0 1/1/1 1/1/1	0/1/0 4/4/1 3/5/1	1/1/1 5/6/1 3/5/4	0/1/0 1/1/0 1/1/1	0/0/0 1/1/0 1/1/1	0/0/0 1/0/0 1/1/1	0/0/0 1/0/0 0/0/0
4	1/1/1 2/3/1 0/1/1	1/3/1 7/7/3 2/9/9	5/12/12 12/12/12 12/12/12	6/12/12 12/12/12 12/12/12	0/1/1 1/1/1 1/1/1	0/1/0 1/1/0 0/1/1	0/0/0 0/3/0 0/1/1	0/1/0 1/1/1 0/0/0
5	1/1/1 1/0/0 1/1/0	1/7/3 4/6/1 4/12/10	9/12/12 12/12/6 12/12/12	3/12/12 9/12/3 9/12/12	0/0/0 0/0/0 0/0/0	0/1/1 1/1/0 3/9/4	0/0/0 1/1/0 2/6/5	0/0/0 0/0/0 0/0/0
6	0/0/1 1/2/0 0/1/0	1/6/1 3/3/1 2/10/5	1/6/4 5/4/3 4/10/11	0/1/0 0/1/1 1/1/1	0/1/0 1/0/0 0/1/1	1/2/1 4/4/0 4/12/10	0/0/0 1/1/0 1/2/1	0/0/0 0/0/0 0/0/0
7	0/1/0 1/4/1 0/1/1	1/4/1 9/9/4 4/12/6	1/4/3 10/12/6 9/12/12	0/0/1 4/5/1 1/4/3	0/0/0 0/0/0 0/0/0	0/2/1 3/1/1 1/6/3	0/1/0 1/1/0 0/0/1	0/0/0 1/1/1 0/0/1
8	3/9/1 10/12/3 0/4/3	4/12/5 12/12/5 8/12/12	5/12/8 12/12/12 11/12/12	4/12/6 12/12/6 9/12/12	0/0/0 1/1/1 1/0/1	0/0/0 1/1/0 1/1/0	0/1/0 0/1/0 1/1/1	0/1/0 0/0/0 0/0/0
	Number of histidine clusters in parses			**60 / 180 / 300**				
	a	***b***	***c***	***D***	***e***	***f***	***g***	***h***
1	9/5/1	12/6/12	9/7/10	3/2/5	0/0/0	1/1/1	0/1/0	0/0/1
2	1/0/0	1/1/3	0/0/2	0/0/0	0/0/0	0/0/3	0/0/0	0/0/0
3	0/0/0	1/3/3	1/2/2	0/4/1	0/1/0	0/1/1	0/0/0	0/0/0
4	1/2/1	1/7/9	10/12/12	11/12/12	0/1/1	0/1/1	0/2/2	0/0/0
5	1/0/2	7/3/12	12/8/12	10/6/8	0/0/0	0/1/9	0/1/7	0/0/0
6	1/1/1	5/3/12	3/4/9	0/1/0	0/0/0	5/8/12	0/0/7	0/0/0
7	1/4/0	4/12/12	1/12/11	1/1/1	1/1/1	2/4/5	0/1/2	0/0/1
8	8/12/3	12/12/12	6/12/12	6/12/10	0/1/1	0/0/1	0/1/3	1/1/0

due to the high presence of ASP residues on protein exteriors. Indeed, many clusters of ASP within our studied chess squares show intense positive and negative polar interactions surrounding the carboxylate, particularly in clusters with low SASA. Those maps that appear largely void of interactions are in clusters with high solvent-accessible surface area, where, as we noted above, there are no residue-protein interactions.

For brevity, we are discussing in more detail ASP residues in the b1 chess square, but further detail on the c5, d5 and f6 chess square results are in Supporting Information. Aspartic acid residues in the b1 chess square appear to be, comparatively, the least solvent-exposed of the four squares, yielding more robust sidechain interactions; this point is the subject of further discussion in a later section. Figures 2.8–2.10 display the contoured maps for ASP in the 60°, 180° and 300° parses of b1, respectively. Additional representative maps from the c5, d5, and f6 chess squares of ASP are visible in Figures 2.11–2.13. The percentile contribution of each cluster to the chess square/parse is listed, along with the average GETAREA[15] SASA (S) and the fraction of the members of that cluster that are protonated (f_{prot}).

One significant point is that the displayed contours, as they represent a map, are showing interactions. Thus, cases where the ASP is ionized (acting as an H-bond acceptor) interacting with a donor could be indistinguishable from cases where the ASP is protonated (acting as a donor) interacting with an acceptor. Thus, it is entirely reasonable for some clusters to have a mixture of ionized and protonated ASPs, although most have $f_{prot} \leq 0.2$ or $f_{prot} \geq 0.8$. Most interactions shown are of the positive polar type, which is appropriate, given the role we expect ASP to serve. These are the prominent, mostly blue

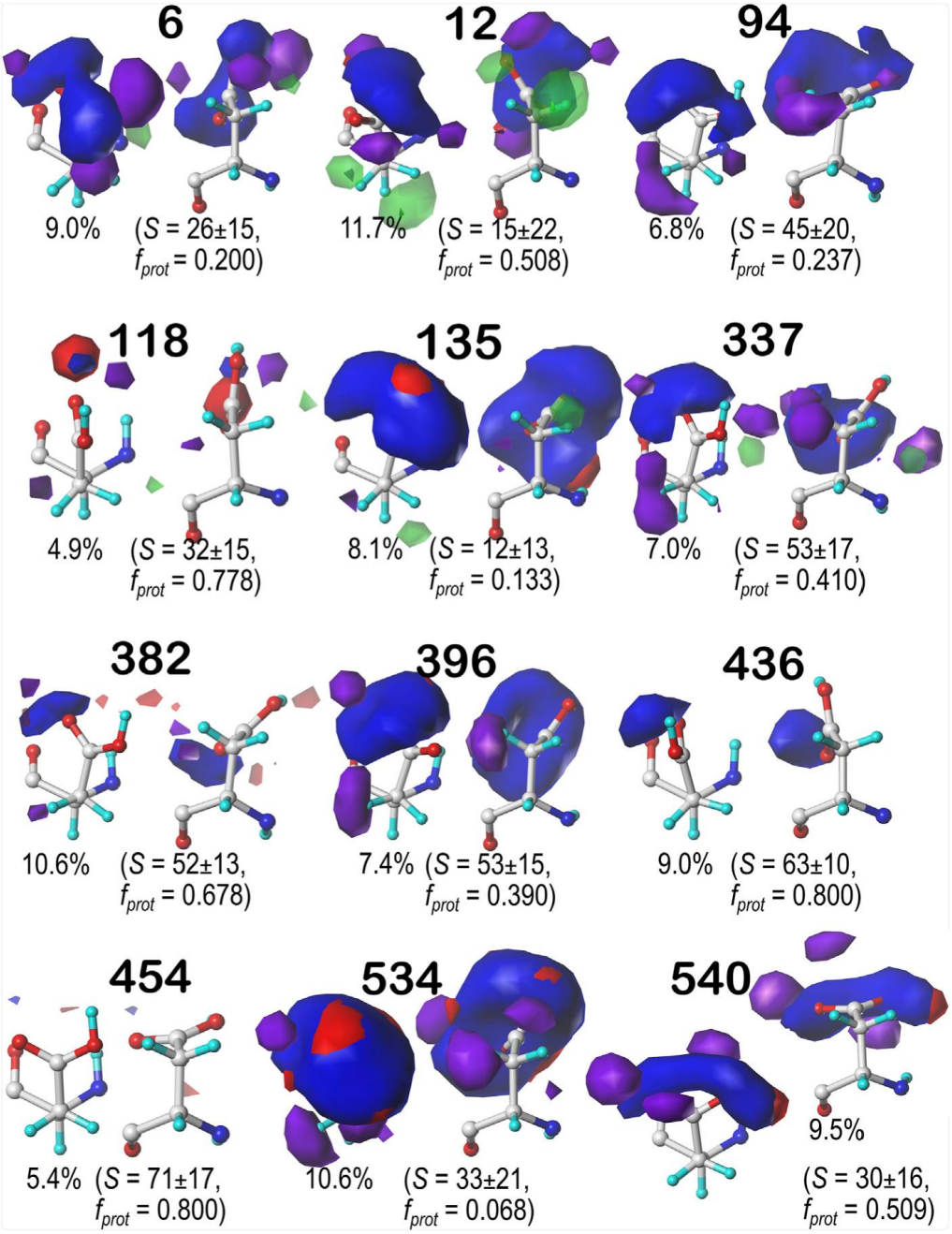


Figure 2.8. Hydropathic interaction maps displaying the Gaussian-weighted average sidechain environments of aspartic acid in the $\chi_1 = 60°$ parse of the *b1* chess square at pH = 3.345. Two map viewpoints are given for each cluster, whose ID is given in bold. The left map in each pair is oriented such that the CA-CB z-axis bond points upward, while the right is oriented to point it out of the page. The x-axis is oriented horizontally in both. The percentage indicates the fraction of the parse represented by that cluster. S represents the solvent accessible surface area in Å2, and f_{prot} indicates the fraction of the cluster protonated at pH50. Blue contours indicate positive polar interactions made with the sidechain, and red indicates negative polar interactions, while green and purple indicate positive and negative hydrophobic interactions, respectively.

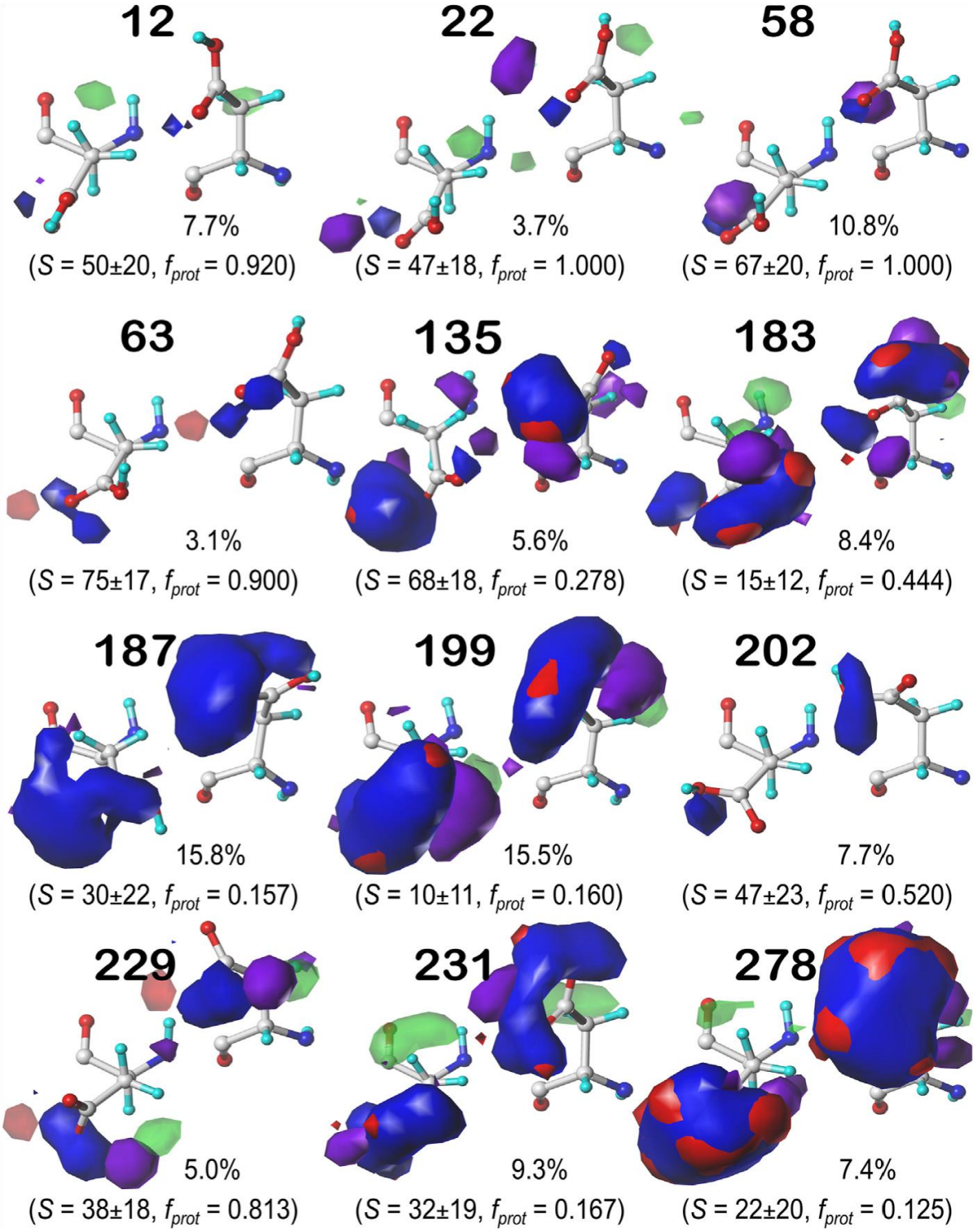


Figure 2.9. Hydropathic interaction maps displaying the Gaussian-weighted average sidechain environments of aspartic acid in the $\chi_1 = 180°$ parse of the b1 chess square at pH = 3.345. See caption for Figure 2.8.

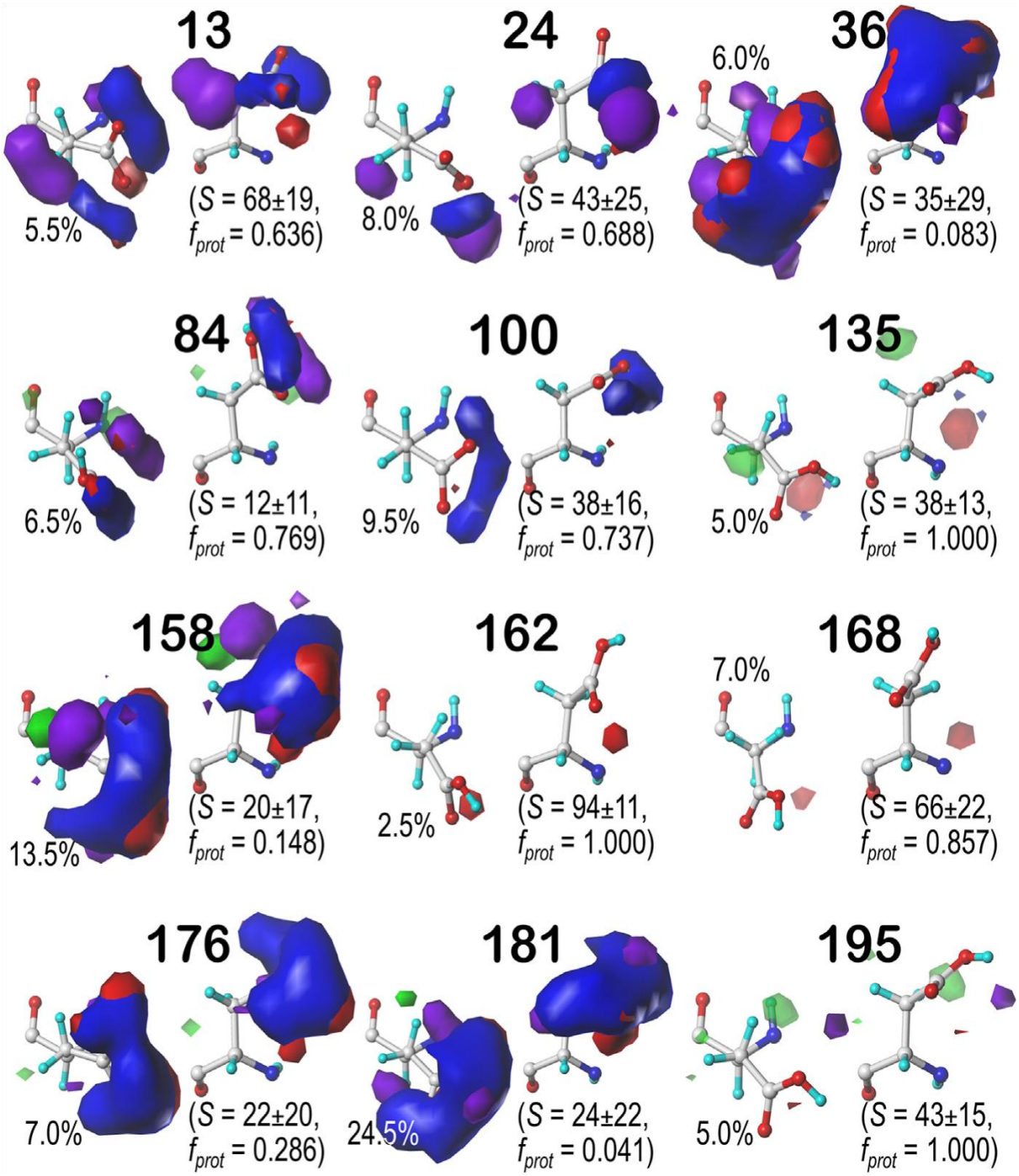


Figure 2.10. Hydropathic interaction maps displaying the Gaussian-weighted average sidechain environments of aspartic acid in the χ1 = 300° parse of the *b1* chess square at pH = 3.345. See caption for Figure 2.8.

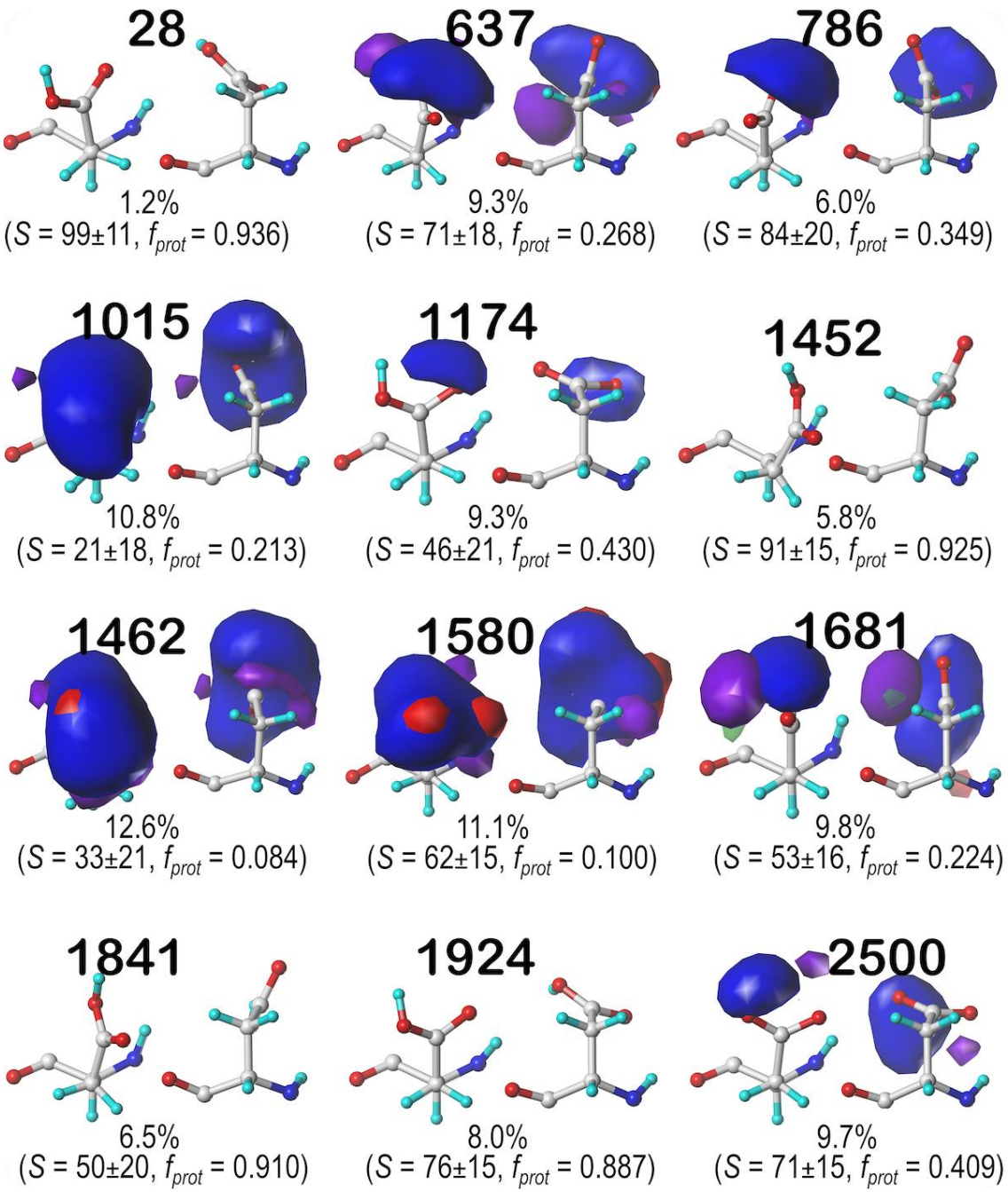


Figure 2.11. Hydropathic interaction maps displaying the Gaussian-weighted average sidechain environments of aspartic acid in the $\chi 1 = 60°$ parse of the *c5* chess square at pH = 3.345. See caption for Figure 2.8.

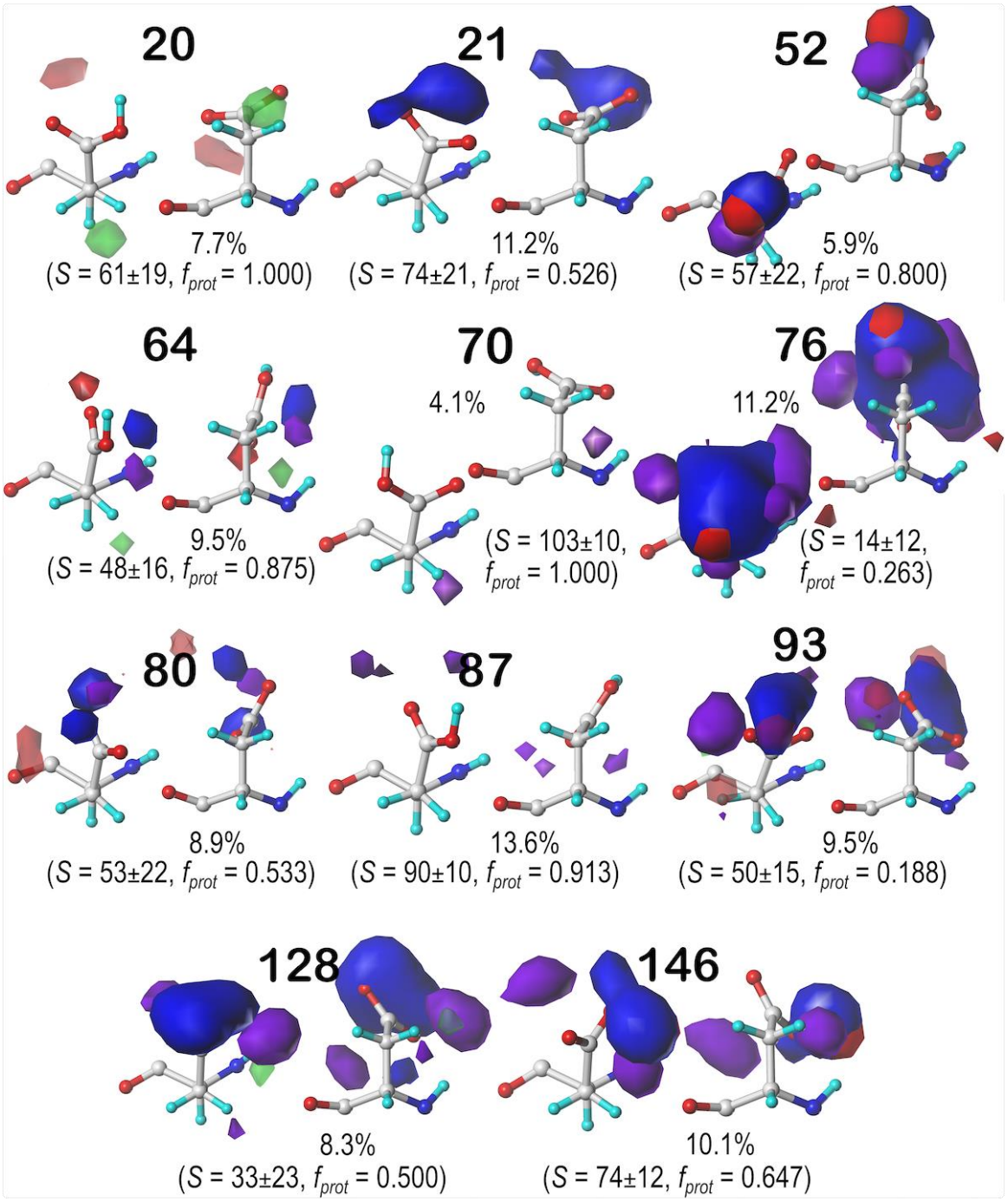


Figure 2 . 12. Hydropathic interaction maps displaying the Gaussian-weighted average sidechain environments of aspartic acid in the $\chi_1 = 60°$ parse of the *d5* chess square at pH = 3.345. See caption for Figure 2.8.

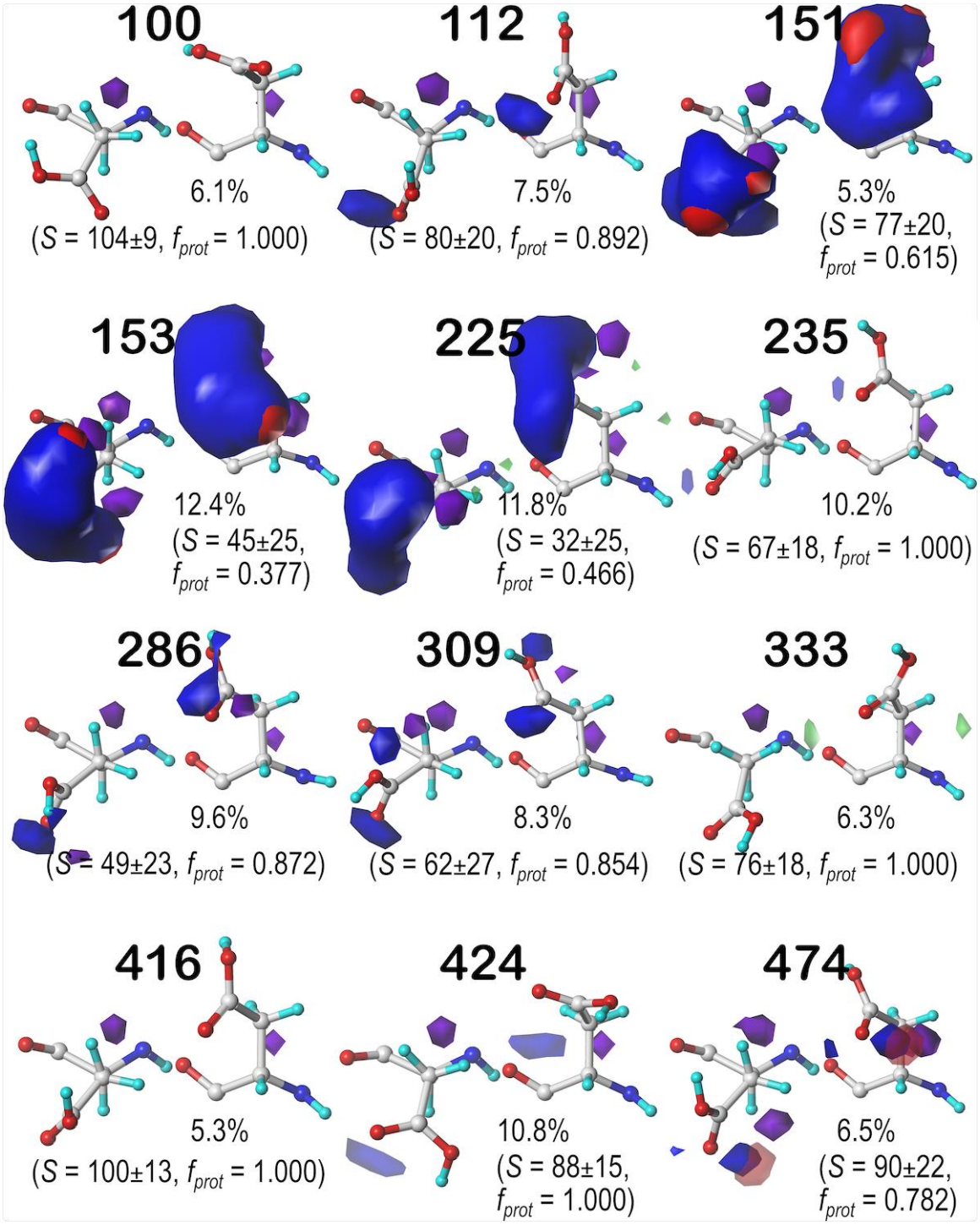


Figure 2 . 13. Hydropathic interaction maps displaying the Gaussian-weighted average sidechain environments of aspartic acid in the $\chi 1 = 180°$ parse of the *f6* chess square at pH = 3.345. See caption for Figure 2.8.

contours near the carboxy acid/carboxylate oxygens that signify hydrogen bonds between one or both of these atoms and their environment. Additionally, many clusters in buried environments with low SASA (<20 $Å^2$) were calculated to be largely deprotonated, i.e., ASP in this environment is acting as a hydrogen bond acceptor. However, some clusters showed high degrees of protonation at pH_{50} = 3.345, such as clusters 12, 118 and 540 in b1.60 (Figure 2.7) and 84 in b1.300 (Figure 2.9). Cluster 84, in particular, showed protonation of 77% of its members with a SASA of 13 ± 12 $Å^2$ at this pH.

Contour maps for the c5, d5 and f6 chess squares show largely similar map profiles and are presented in Supplementary Figures S2 and S3 for c5 parses 0.180 and 0.300, respectively; in Supplementary Figures S5 and S6 for d5 parses 0.180 and 0.300, respectively; and in Supplementary Figures S7 and S9[1] for f6 parses 0.60 and 0.300, respectively. Further numerical data supporting these results and encompassing all chess squares is provided in Supplementary Table S5.[1] In summary, each map appears to be a backbone-specific representation of a unique collection of interactions made by an aspartate/aspartic acid residue. To demonstrate this, we calculated inter-cluster similarities using the previously described algorithms. The average cluster-cluster similarities within chess squares are: 0.799 in b1, 0.795 in c5, 0.791 in d5, and 0.802 in f6 chess squares. However, a few pairs of cluster maps in the adjacent chess squares c5 and d5 have similarities of >0.900: 637 (c5.60) and 146 (d5.60), 57 (c5.180) and 70 (d5.180), and 217 (c5.300) and 58 (d5.300), indicating that backbone secondary structural elements may encode inherent similarities in the kinds of environments likely to surround a given residue.

Glutamic Acid

Glutamic acid tells a very similar story to that of aspartic acid, so many of the points made for that residue stand here, as well. First, the bulk of interactions made with the GLU sidechain are of the positive polar type, followed by negative polar. Again, many clusters were also calculated to have high SASA. Also, we calculated GLU maps with three times as many parses as ASP (vide supra), due to the 1-carbon extension to its sidechain, making the number of clusters about three times as many. We believed it is redundant to showcase maps for every average cluster in every subparse. Instead, we have chosen to focus on the b1 chess square and show maps of its highest occupied clusters in each parse (Figure 2.14). This collection is representative of the 67 b1 clusters suggests the diversity of sidechain orientations available in the full map set. One aspect of the GLU maps that we expected to see was an amplified presence of hydrophobic interactions compared to the ASP maps. However slightly, the maps of these specific clusters do show some indication of additional hydrophobic interactions localized around the hydrophobic chain, although these interactions appear more likely in the lower population parses.

Their lack of visibility in Figure 2.10 may be more due to the limitations of contouring at consistent values than anything else, but perhaps the expected hydrophobic interactions with this sidechain are actually rare or have backbone conformation dependence. A confounding factor certainly is that GLU is even more solvent exposed than ASP, and this will be explored below. Numerical data for all GLU chess squares is provided in Supplementary Table S6.[1]

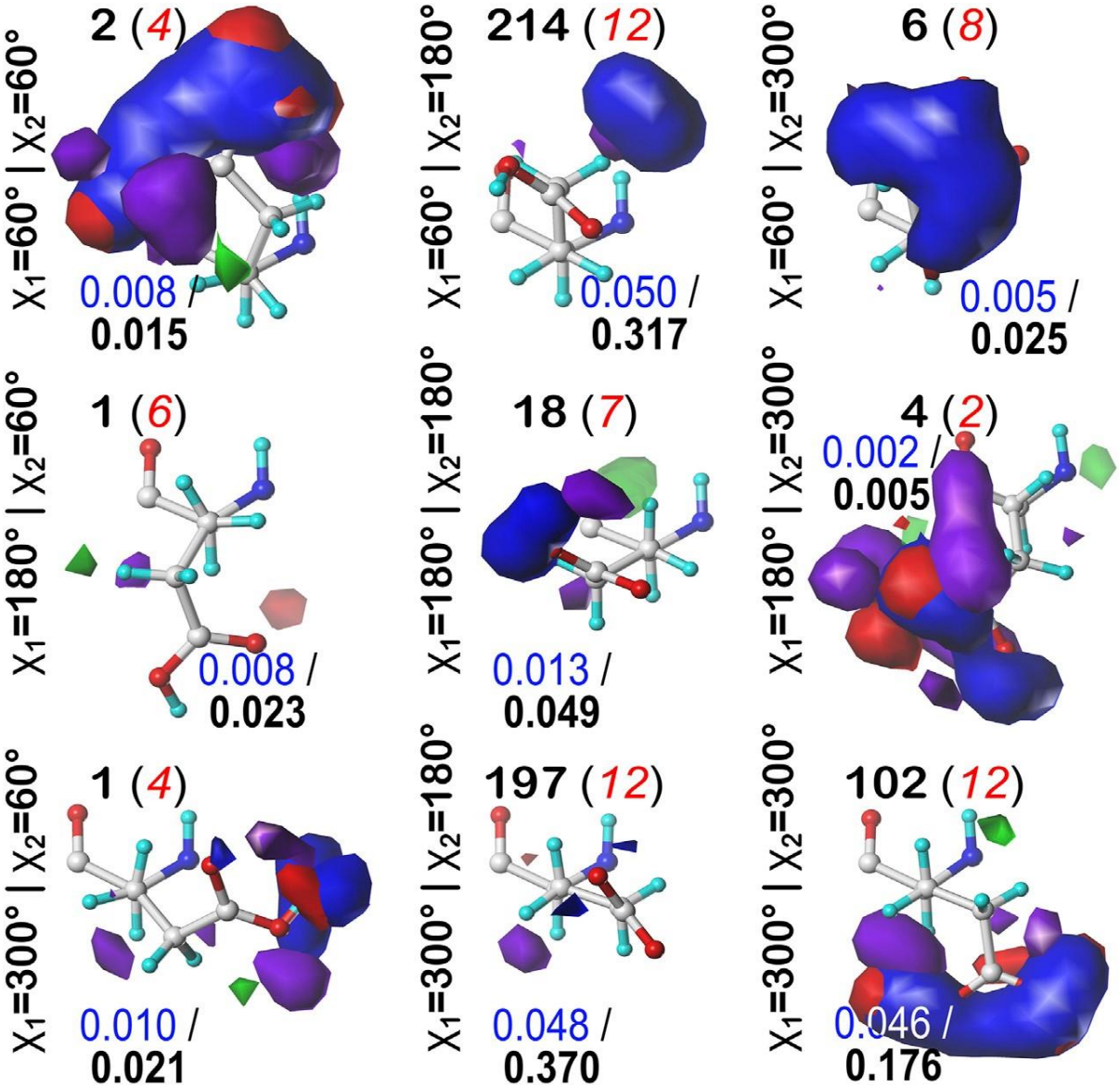


Figure 2.14. Hydropathic interaction maps displaying the Gaussian-weighted average sidechain environments of glutamic acid in the highest populated clusters of the nine parses of the *b1* chess square at pH = 4.224. Residues are oriented such that the CA-CB z-axis points upward and the x-axis runs to the right. The parses of the χ_1 and χ_2 angles are indicated along the side of each map. The cluster ID and number of clusters in the parse are given above the map in black and red, respectively. Below each map, in blue, is indicated the fraction of the entire chess square represented by each map, followed in black by the parse's representative fraction of the chess square. Blue contours indicate position and magnitude of positive polar interactions near the sidechain, while red represents negative polar interactions. Green and purple contours indicate positive and negative hydrophobic interactions, respectively.

Histidine

Histidine naturally tells very much a different story from ASP and GLU. Its imidazole sidechain can play numerous roles in protein structure. Not only does it have more protonation states than the acidic residues we have discussed, but its two nitrogens can act as either (or both) hydrogen bond donors and acceptors in any combination. Its ring is partially hydrophobic and aromatic, meaning it can make any variety of polar, nonpolar, and π-π stacking interactions with other residues. These π-π stacking interactions with aromatic residues, for example, may be indicated in maps where the ring is bordered by large, flat, green contours. This brand of versatility is very clearly indicated in our generated maps for HIS. Figure 2.15 displays the contour maps for the HIS b1.60 chess square parse. Figures 2.16 through 2.18 show exemplary histidine maps in the c5, d5 and f6 chess squares. The patterns in these maps are complex, but interpretable in terms of the interaction types. A detailed description for all 12 clustered maps in the 0.60 parse of the b1 chess square would be too much for here, but first, it is clear that all maps displayed here (and in Supplementary Figures S20–S30)[1] represent unique sets of interaction features, or routes to complete the residue's hydropathic valences. Consider cluster 31 in the b1.60 map set (Figure 2.15): 93.3% of the histidines in this cluster are protonated, it has mid-range solvent exposure, the CB methylene is making hydrophobic interactions (green) with its environment, and the protonated NE is engaged in a hydrogen bonding interaction (blue) largely perpendicular to the ring. Cluster 235 here is singly protonated at NE, which engages with an on-axis hydrogen bond, and has very low solvent exposure, and its environment is dominated by hydrophobic interactions,

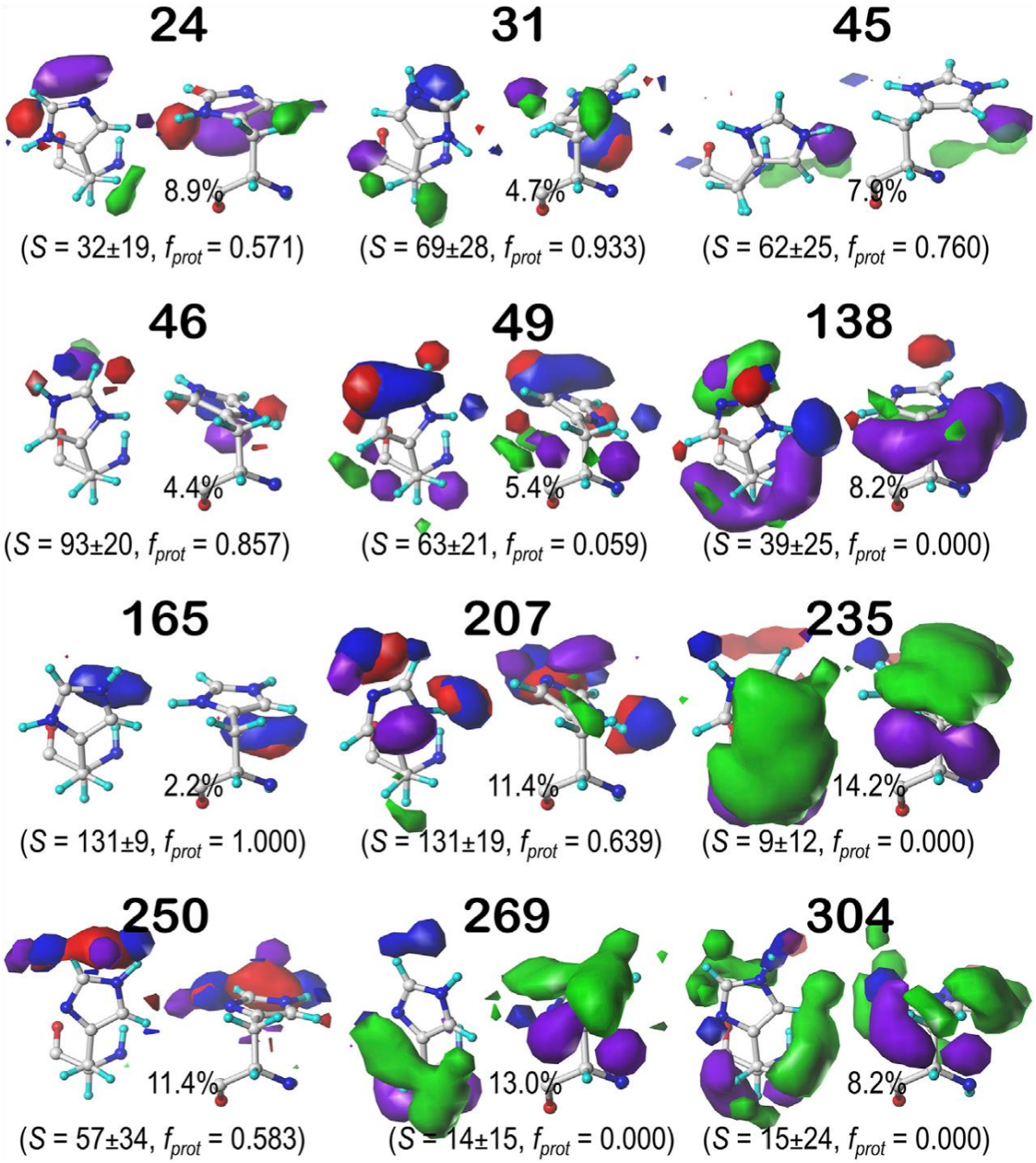


Figure 2.15. Hydropathic interaction maps displaying the Gaussian-weighted average sidechain environments of histidine in the $\chi_1 = 60°$ parse of the *b1* chess square at pH = 5.174. See caption for Figure 2.8.

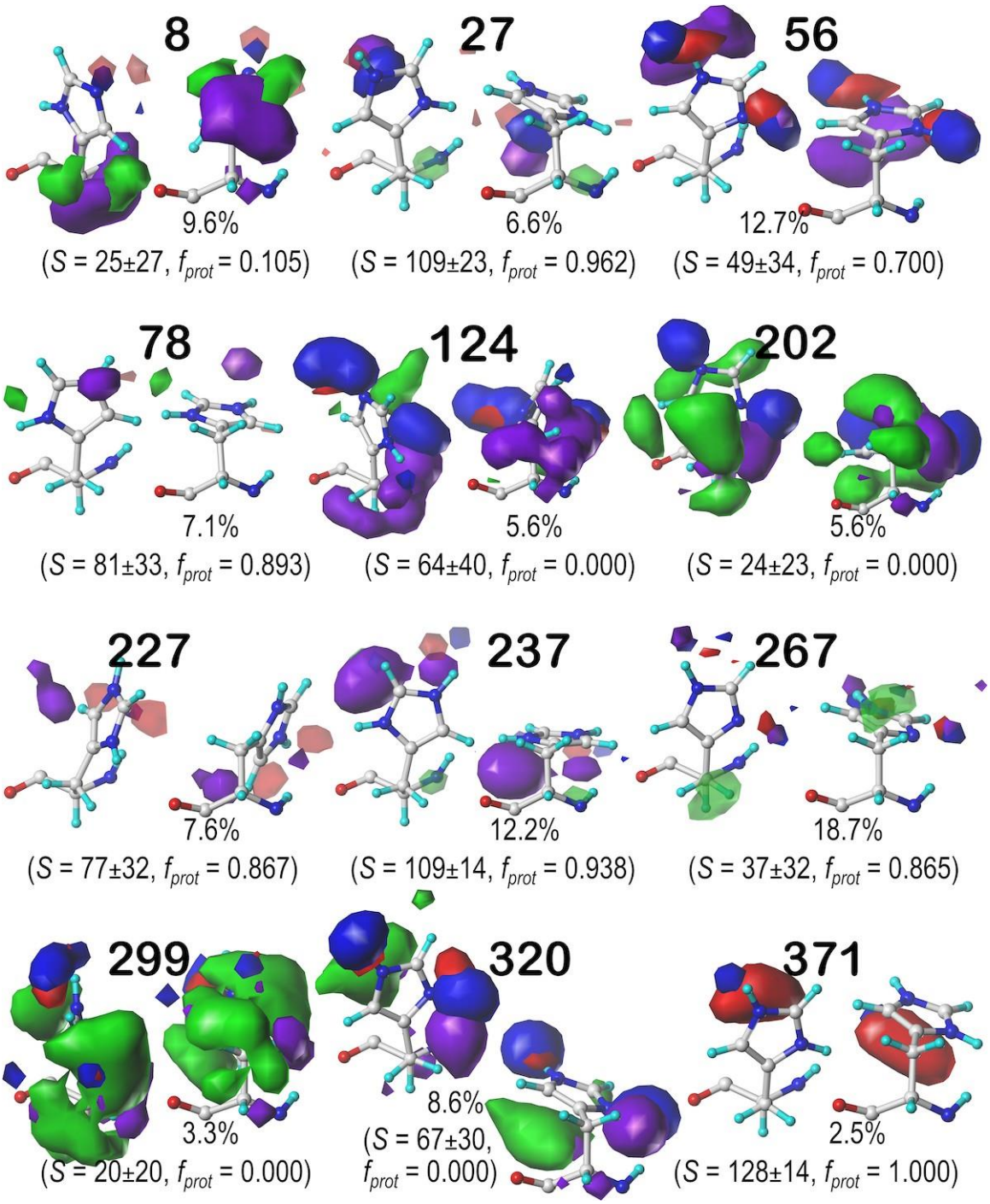


Figure 2.16. Hydropathic interaction maps displaying the Gaussian-weighted average sidechain environments of histidine in the χ_1 = 60° parse of the *c5* chess square at pH = 5.174. See caption for Figure 2.8.

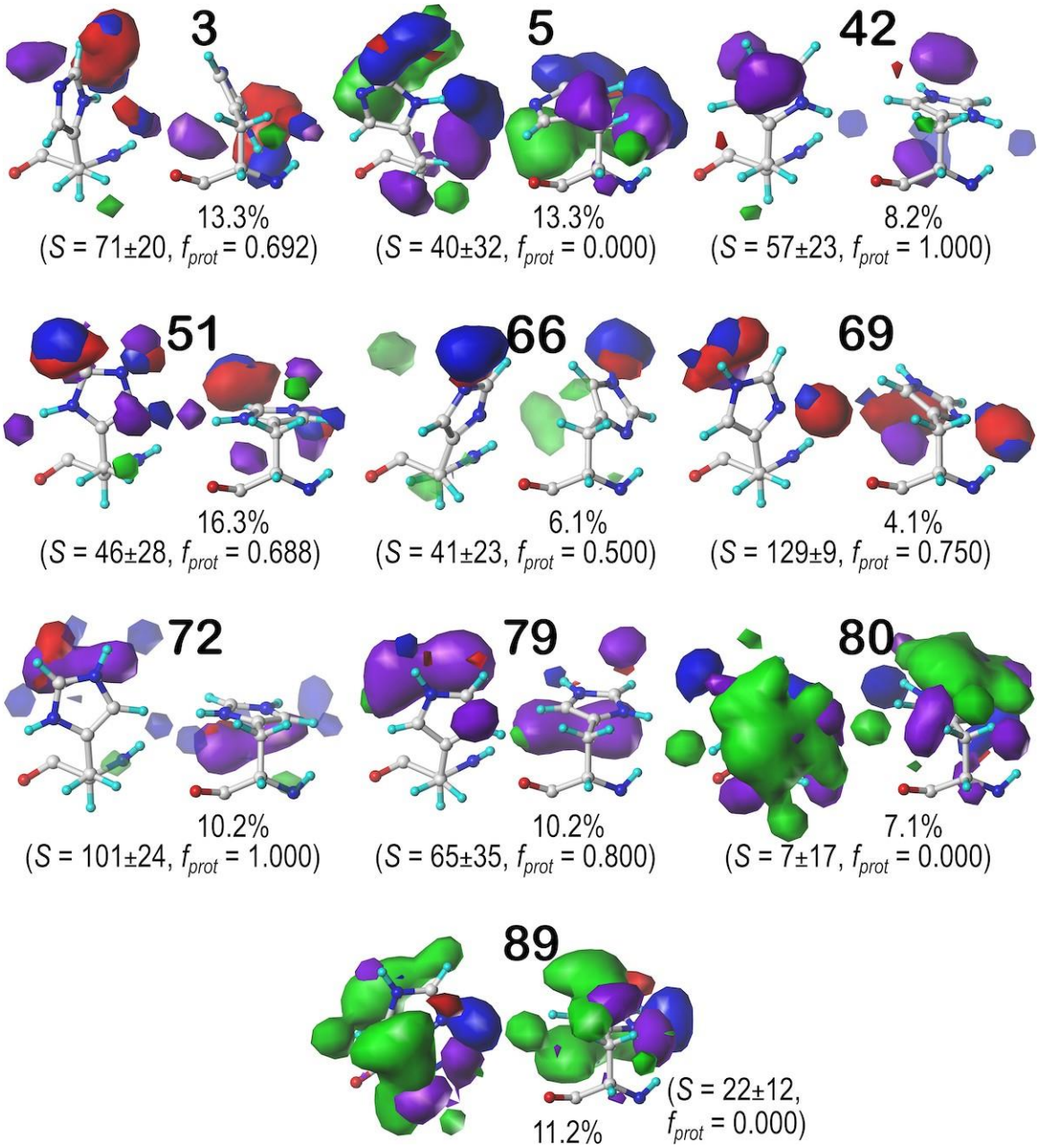


Figure 2 . 17. Hydropathic interaction maps displaying the Gaussian-weighted average sidechain environments of histidine in the χ_1 = 60° parse of the *d5* chess square at pH = 5.174. See caption for Figure 2.8.

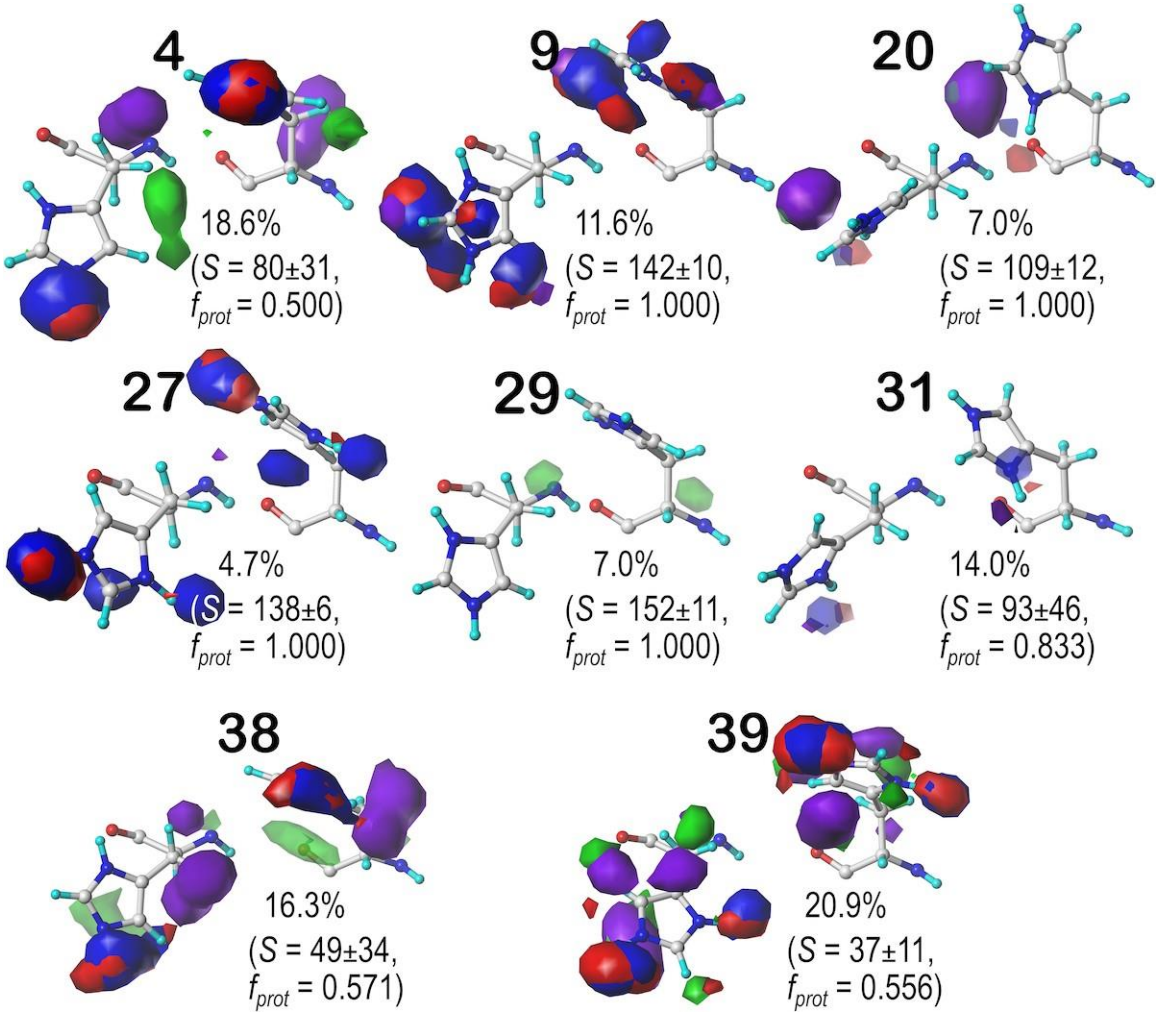


Figure 2.18. Hydropathic interaction maps displaying the Gaussian-weighted average sidechain environments of histidine in the χ_1 = 180° parse of the *f6* chess square at pH = 5.174. See caption for Figure 2.8.

both favorable (green) and unfavorable (purple), with the former above the ring and the latter below the ring. Comprehensive numerical data for all chess squares of histidine is provided in Supplementary Table S7.[1]

Hydropathic Character of Maps With Changes in pH

We were interested to see how changing the environmental pH would affect the maps. In other words, can we rationally "tune" the residue interactions by this means, and

can that be exploited in protein design, e.g., to stabilize or destabilize binding sites, folds or interfaces? As an illustration, consider ASP141A in PDB structure 1WNS—family B DNA polymerase from hyperthermophilic archaeon pyrococcus kodakaraensis KOD1,[26] which is situated in a highly anionic region with three other acidic residue side chains. This residue is in our cluster 202 of parse b1.180 with f_{prot} = 0.520 and has a significant free energy difference between protonated and deprotonated states. Our model suggests ASP141A has an elevated pK_a and, when protonated, forms a hydrogen bond with ASP215A. There are significant visible differences between the calculated maps for this particular residue (Figure 2.19): at high pH (9), the interactions surrounding ASP141A (top) are largely unfavorable polar, but protonation, as shown in the low pH (5) case, protonates one of the carboxylate oxygens and yields a strong favorable hydrogen bond between it and ASP215A. As described earlier, the map contours displayed in this work were calculated at what we are calling pH_{50}, which shows the highest diversity of protonated and deprotonated cases. Such maps can be calculated, clustered, etc. at any pH, and indeed making use of different maps at different protonation states will expand the scope for protein structure prediction of real situations where ionization states can vary due to local environments.

For further insight, we examined the interaction character of ASPs in one parse, b1.300, to determine if the relative fractions of our four-type quartet of interactions were altered with changes in pH (Figure 2.20). We expected to see small, but noticeable, changes in clustering of residues as adjustment of pH altered the memberships of the clusters as protonation became either more favorable or unfavorable. To facilitate

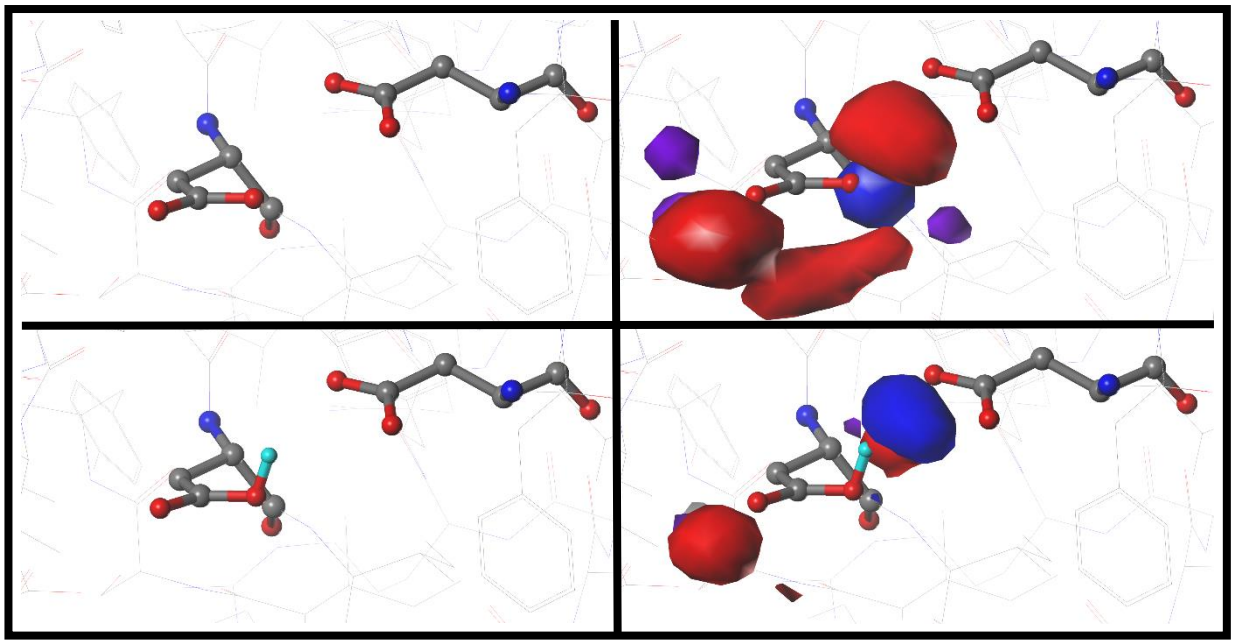


Figure 2.19. Variations in mapped environments around ASP141A in PDB structure 1WNS. A) structure model mapped environment around deprotonated ASP141A with strong unfavorable polar interaction between it and nearby residue ASP215A (pH 9). B) structure model and mapped environment around protonated ASP141A with new strong, favorable polar interaction with ASP215A (pH 5).

comparisons between the cluster sets at different pH values, the bars are arranged by increasing average solvent-accessible surface area for the cluster (low to high). At pHs of 1, 3.345 (i.e., pH_{50}) and 7, some character changes were in fact observed, but, interestingly, most of these occurred in low population clusters. We theorize that, as residues clustered differently, residues being added/subtracted to/from new groups simply had a greater impact on the overall character of smaller clusters. One point of note, however, is that, although most clusters with high SASA had the highest protonation levels (discussed later), only cluster 84 retained any level of protonation at pH 7, in spite of having the lowest SASA. This suggests that this cluster, in particular, describes scenarios where aspartate protonation is energetically required.

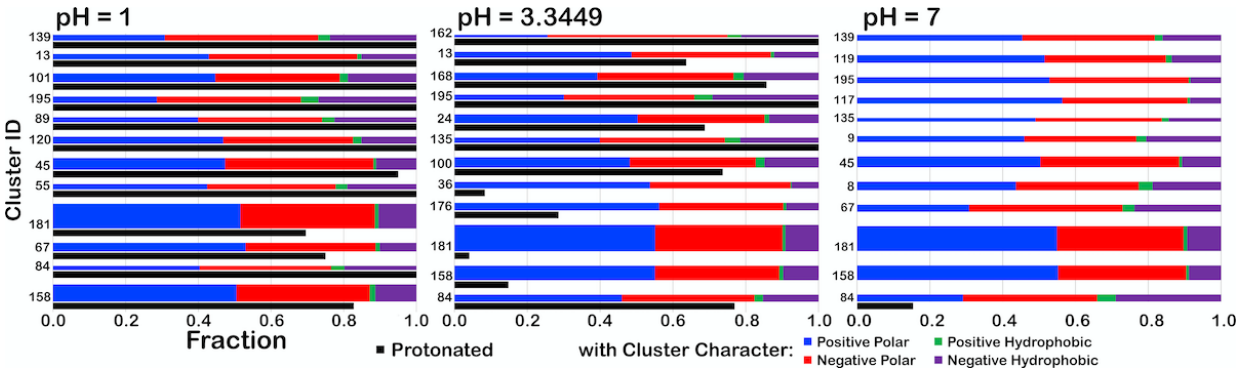


Figure 2.20. Character interaction charts for ASP residues in the b1.300 parse at pH 1, 3.345, and 7. The fraction of each interaction type is given on the x-axis, for each cluster ID on the y-axis. The bars are arranged such that, descending, clusters have smaller SASAs. The thickness of the bars indicates residue population contained within that cluster. The black bars indicate f_{prot}, the fraction of the residues in the cluster protonated.

We also examined the interaction character of the GLU b1.300.180 parse (Figure 2.21), which is probably the parse most like the b1.300 parse of ASP. The clusters within this GLU parse generally involved more hydrophobic interactions, both favorable and unfavorable, than those of the ASP b1.300 parse. However, these observations are subtle and not easily visualized in the map contours. Nevertheless, overall, the average fractions of favorable and unfavorable hydrophobic interaction contributions, $f_{hydro(+)}$ and $f_{hydro(-)}$, are 0.038 and 0.218, respectively for GLU, and 0.021 and 0.153 for ASP at their respective pH_{50}s. Importantly, the higher propensity for hydrophobic interactions by GLU, due to the additional methylene in the sidechain, are encoded in the interaction maps on a cluster by cluster basis.

Our ability to generate tunable maps for HIS is slightly more limited. The constrained conformational flexibility of the HIS sidechain and surrounding protein allowed by our approach could clearly be remedied by molecular dynamics or even energy minimization, but the cost–beyond CPU, etc. –would be the loss of positional certainty afforded by experimental data. That said, our map data for HIS, like ASP and

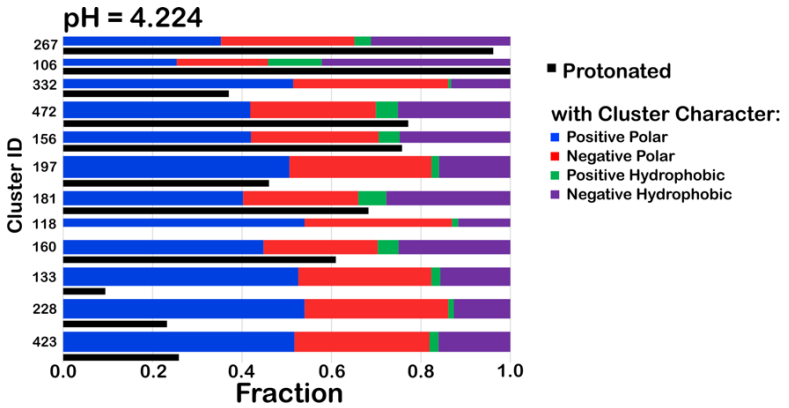


Figure 2.21. Character interaction chart for the GLU *b1.300.180* parse at pH 4.224. The fraction of each interaction type is given on the x-axis, for each cluster ID on the y-axis. The bars are arranged such that, descending, clusters have smaller SASAs. The thickness of the bars indicates residue population contained within that cluster. The black bars indicate f_{prot}, the fraction of the residues in the cluster protonated.

GLU, exhaustively captures the many possible HIS interaction environments found in crystallographic structures exploitable for protein structure analyses and predictions.

Solvent-Accessible Surface Areas for the Ionizable Residues

The historical Ramachandran plots showed the relationship between backbone angles and frequency of observation. Our chessboard schema (Figure 2.1 for ASP, Figure 2.15 for GLU and HIS) was intended to organize our dataset by backbone structure, and thus facilitate comparisons between like residues. We also see a further population dependence on χ1 (and χ2 for GLU). In fact, further exploration revealed that solvent accessibility for each of our three residues is also seemingly dependent on the residue's backbone and χ angles, which suggests a trend between this level of solvent exposure and underlying protein structure. For example, the average SASAs for ASP residues were calculated to be 37, 59, 64, and 64 Å^2 for the b1, c5, d5, and f6 chess squares, respectively. With a similar trend, the average SASAs for GLU residues were calculated to be 57, 75, 80, and 81 Å^2 for the b1, c5, d5, and f6 chess squares, respectively. However, in spite of it being significantly more hydrophobic than ASP and GLU, and thus more likely to be

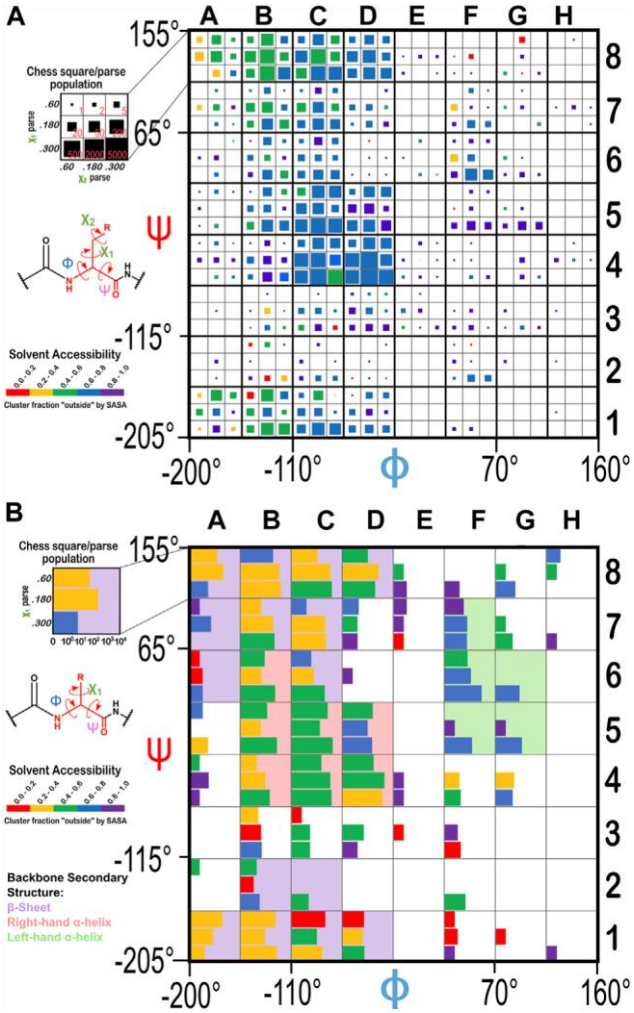


Figure 2.22. Ramachandran chessboard displaying the chess square/parse population for A) glutamic acid and B) histidine. The (χ1/χ2) parse populations for GLU are represented by colored squares with sizes as indicated on the legend. The (χ1) parse populations for HIS are represented in log10 scale with colored bars. See also caption for Figure 2.1.

buried, GETAREA calculations for HIS yielded the surprisingly large average SASAs of 41, 59, 62, and 79 Å^2 for the b1, c5, d5, f6 chess squares, respectively.

To evaluate our data in a more nuanced way, we calculated the "fraction outside" ($f_{outside}$) metric based on GETAREA,[35] as described in Methods. The $f_{outside}$ values for each chess square/parse are also illustrated in Figures 2.3 and 2.22, with the colors of the bars (that represent parse populations by their lengths) for ASP and HIS or squares (that represent parse populations by their areas) for GLU. Chess square/parses within the β-pleat region of the Ramachandran plot for aspartate (Figure 2.3), as expected, show lower $f_{outside}$ (more buried) relative to the right- and left-hand α-helix, i.e., most parses show averaged $f_{outside}$ in the 0.4–0.6 (green) range, whereas in the α-helix region most are in the $f_{outside}$ range 0.6–0.8, and the left-hand α-helix is still more exposed, in the $f_{outside}$ range 0.8–1. 0. The same trends hold for glutamates (Figure 2.22A), although the data suggests somewhat larger $f_{outside}$ values. This is likely a result of GLU's inherent

additional surface area concomitant with its 1-carbon chain extension. The $f_{outside}$ trends for HIS (Figure 2.22B) suggest more buriedness: in the β-pleat region of the Ramachandran plot, the parses are evenly split between the 0.2–0.4 and 0.4–0.6 ranges (yellow and green), histidines in the α-helix region are in the $f_{outside}$ range 0.4–0.6, while those in the left-hand α-helix are more exposed, in the range 0.6–0.8.

It should be noted that the sidechain solvent-accessible surface areas for these three residues in Gly-X-Gly "random coil" tripeptides show that histidine has a larger surface area (154.6 $Å^2$) than either aspartate (113.0 $Å^2$) or glutamate (141.2 $Å^2$),[15] which is incorporated into the $f_{outside}$ calculations. Thus, while HIS may have, overall, higher solvent exposure in surface area, the actual fraction of solvent-exposed residues is smaller. All three residues show the same trend: larger solvent exposure in the α-helix regions that is more extreme in the left-hand region, and greater burial in the β-pleat region. These conclusions are in qualitative agreement with those of Lins et al.[27] in their report on differences in solvent-accessible surface area between residues in different secondary structures. However, $f_{outside}$, exactly as SASA does, varies from cluster-to-cluster within each chess square and parse. For example, $f_{outside}$ for ASP b1.300 ranges widely–between 0.077 (cluster 84) to 1.000 (cluster 162), despite its overall $f_{outside}$ of <0.4 suggesting mostly burial for this group of residues. The SASA and $f_{outside}$ values for all three residues in this study, on a cluster-by-cluster basis are included in the Supplementary Tables S5–S7. To summarize, each 3D map cluster represents a unique set of interactions that also encodes solvent exposure and buriedness. We should emphasize that map profiles appearing to be similar could manifest with different buriedness and/or protonation, and thus remain unique.

Summary and Conclusion

We analyzed the interaction environments of more than 105,000 ionizable amino acid residues (aspartic acid, glutamic acid, histidine) in a diverse collection of protein structures. From above and our previous reports,[7,8] it is clear that the hydropathic environment surrounding an amino acid residue in a protein can be mapped in terms of its interactions. Significantly, the patterns of interactions within the maps, representing the constellation of contacts and their interaction strengths and characters, cluster into a fairly limited set of unique, backbone-dependent motifs. Each of these motifs can be rendered into an average map quartet and an average prototype residue structure. Thus, we have produced a backbone-dependent library of not only sidechain rotamers, but also 3D residue interaction preferences. The presence of a feature, such as a favorable polar interaction in one of these maps, e.g., an ASP in the b1.300 (β-pleat) cluster 100 (Figure 9), where the carboxylate/carboxylic acid functional group is involved in hydrogen bonding through both oxygens, should have complementary donors/acceptors on neighboring residue(s). Accordingly, those residue's maps should contain similar features, and the alignment of these features–and all others from a collection of such maps–would describe a well-organized hydropathic interaction network.

It is not just the favorable hydrophobic and polar interactions that constitute this network. The maps illustrated by contours here, and previously,[7-9] nearly ubiquitously display unfavorable polar and hydrophobic interactions. These interactions are integral parts of protein structure; for example, even polar residues like the ASP, GLU, and HIS of this report have hydrophobic atoms covalently bonded to the polar functional groups.

Thus, a background of unfavorable hydrophobic interactions is usually seen with strong favorable polar interactions. However, other hydrophobic interactions are functional components of structure that Nature uses, e.g., for adding flexibility or isolating water. Developing an understanding of them will help illuminate protein design and drug discovery. Unfavorable polar interactions, on the other hand, provide a route to understanding and predicting residue ionization states. The presence of this type of interaction signals an opportunity for water intervention, an adjustment in local pH or can be used as drug design cues.

While our predictions of pK_as for ASP and GLU are adequate (and seemingly less so for HIS over a much smaller training set), our primary goal was not that, but instead to evaluate the hydropathic environments surrounding these residue types. As expected, those environments change drastically with pH. We illustrated environments with 3D maps for an artificial halfway point–pH_{50}–that showed a range of environments, but we have also calculated maps for other pH cases, and the nature of interactions displayed therein are, although unsurprising, quite informative. Importantly, this means that we can tune residue hydropathic environment maps as a function of pH, and that they encode this critical element of structure, interaction, and energetics in a rational way. Thus, if we use these maps as part of a scheme for protein structure building and prediction, we have the additional scope to explore ionization states in understanding and defining optimal protein structures.

In our 2019 report,[8] we stated that full understanding of the individual environment maps for alanine would first require completing the analysis for all residue types. This current report is a status update on that task–for ASP, GLU and HIS. The

remaining residues are in various stages of completion and analysis, and we anticipate additional communications in the near future.

As with alanine, our evaluation of interactions of the ionizable residues with 3D maps backs our interaction homology paradigm–for understanding and potentially predicting protein structure. The hydropathic valence for ASP and GLU is largely satisfied by a functional group that complements the carboxy acid, and some involvement with the CB, CG (and for GLU, the CD) methylenes by a hydrophobic interaction partner, except if the sidechain is fully solvent exposed. HIS is, however, much more complex, involving additional terms such as hydrophobic interactions with aromatic carbons that may be of π-π character and polar interactions that include hydrogen bonding with its ND1 and/or NE2, as either acceptors or donors. As these effects are recorded within the maps, we see that it is the hydropathic "field" of the atoms surrounding a residue, not specific residue types or atoms, that directs its conformation or other properties, including rotameric and secondary structure. Finally, biological structure is a puzzle consisting of a delicate balance of effects, mostly favorable but others seemingly counterproductive. Assembling structure by homology modeling[28-30] or even de novo structure prediction[31-33] involves many puzzle pieces and interactions, but some key information involving, e.g., hydrophobic interactions or residue ionizations is not utilized in the usual Newtonian physics-based approaches.

Our ability to map interactions in 3D space, including a rational means to explore the local pH of individual residues in more or less real time should be advantageous in later studies. Since the maps highlight interactions, building structural models that optimize the map-map overlaps of interactions arising from adjacent or through-space

residue map pairs (or larger sets) could yield a very useful and unique target function for protein structure prediction, likely quite amenable for machine learning optimization.

References

1. Herrington, N. B.; Kellogg, G. E. 3D Interaction Homology: Computational Titration of Aspartic Acid, Glutamic Acid, and Histidine Can Create pH-Tunable Hydropathic Environment Maps. *Front. Mol. Biosci.* **2021**, *8*, 773385.
2. Kellogg, G. E.; Semus, S. F.; Abraham, D. J. HINT: A New Method of Empirical Hydrophobic Field Calculation for CoMFA. *J. Computer-aided Mol. Des*. **1991**, *5*, 545–552. doi:10.1007/BF00135313
3. Kellogg, G. E.; Abraham, D. J. Hydrophobicity: Is LogP(o/w) More Than the Sum of its Parts? *Eur. J. Med. Chem.* **2000**, *35*, 651–661. doi:10.1016/s02235234(00)00167-7
4. Sarkar, A.; Kellogg, G. Hydrophobicity - Shake Flasks, Protein Folding and Drug Discovery. *Curr. Med. Chem.* **2010**, *10*, 67–83. doi:10.2174/ 156802610790232233
5. Fornabaio, M.; Cozzini, P.; Mozzarelli, A.; Abraham, D. J.; Kellogg, G. E. Simple, Intuitive Calculations of Free Energy of Binding for Protein–Ligand Complexes. 2. Computational Titration and pH Effects in Molecular Models of Neuraminidase–Inhibitor Complexes. *J. Med. Chem*. **2003**, *46*, 4487–4500. doi:10.1021/jm0302593
6. Spyrakis, F.; Fornabaio, M.; Cozzini, P.; Mozzarelli, A.; Abraham, D. J.; Kellogg, G. E. Computational Titration Analysis of a Multiprotic HIV-1 Protease–Ligand Complex. *J. Am. Chem. Soc*. **2004**, *126*, 11764–11765. doi:10.1021/ja0465754

7. Ahmed, M. H.; Koparde, V. N.; Safo, M. K.; Neel Scarsdale, J.; Kellogg, G. E. 3D Interaction Homology: The Structurally Known Rotamers of Tyrosine Derive from a Surprisingly Limited Set of Information-Rich Hydropathic Interaction Environments Described by Maps. *Proteins* **2015**, *83*, 1118–1136. doi:10.1002/prot.24813
8. Ahmed, M. H.; Catalano, C.; Portillo, S. C.; Safo, M. K.; Neel Scarsdale, J.; Kellogg, G. E. 3D Interaction Homology: The Hydropathic Interaction Environments of Even Alanine Are Diverse and Provide Novel Structural Insight. *J. Struct. Biol.* **2019**, *207*, 183–198. doi:10.1016/j.jsb.2019.05.007
9. AL Mughram, M. H.; Catalano, C.; Bowry, J. P.; Safo, M. K.; Scarsdale, J. N.; Kellogg, G. E. 3D Interaction Homology: Hydropathic Analyses of the "π-Cation" and "π-π" Interaction Motifs in Phenylalanine, Tyrosine, and Tryptophan Residues. *J. Chem. Inf. Model*. **2021**, *61*, 2937–2956. doi:10.1021/acs.jcim.1c00235
10. Catalano, C.; AL Mughram, M. H.; Guo, Y.; Kellogg, G. E. 3D Interaction Homology: Hydropathic Interaction Environments of Serine and Cysteine Are Strikingly Different and Their Roles Adapt in Membrane Proteins. *Curr. Res. Struct. Biol.* **2021**, *3*, 239–256. doi:10.1016/j.crstbi.2021.09.002
11. Kellogg, G. E.; Fornabaio, M.; Spyrakis, F.; Lodola, A.; Cozzini, P.; Mozzarelli, A.; Abraham, D. J. Getting it Right: Modeling of pH, Solvent and "nearly" Everything Else in Virtual Screening of Biological Targets. *J. Mol. Graphics Model.* **2004**, *22*, 479–486. doi:10.1016/ j.jmgm.2004.03.008
12. George, P.; Hanania, G. I. H.; Irvine, D. H.; Abu-Issa, I. 1090. The Effect of Co-ordination on Ionization. Part IV. Imidazole and its Ferrimyoglobin Complex. *J. Chem. Soc.*, **1964**, 5689–5694. doi:10.1039/JR9640005689

13. Pahari, S.; Sun, L.; Alexov, E. PKAD: a Database of Experimentally Measured pKa Values of Ionizable Groups in Proteins. *Database*, **2019**, *2019*, baz024. doi:10.1093/database/baz024
14. R Core Team. R: A Language and Environment for Statistical Computing. *R Foundation for Statistical Computing.* **2013**, http://www.R-project.org/.
15. Fraczkiewicz, R., Braun, W. Exact and Efficient Analytical Calculation of the Accessible Surface Areas and Their Gradients for Macromolecules. *J. Comput. Chem*. **1998**, *19*, 319–333. doi:10.1002/(sici)1096-987x(199802)19:3<319:aid-jcc6>3.0.co;2-w
16. Ramachandran, G. N.; Ramakrishnan, C.; Sasisekharan, V. Stereochemistry of Polypeptide Chain Configurations. *J. Mol. Biol*. **1963**, *7*, 95–99. doi:10.1016/s0022-2836(63)80023-6
17. Shapovalov, M. V.; Dunbrack, R. L., Jr. A Smoothed Backbone-Dependent Rotamer Library for Proteins Derived from Adaptive Kernel Density Estimates and Regressions. *Structure* **2011**, *19*, 844–858. doi:10.1016/j.str.2011.03.019
18. Harms, M. J.; Castañeda, C. A.; Schlessman, J. L.; Sue, G. R.; Isom, D. G.; Cannon, B. R.; García-Moreno E., B. The pK_a Values of Acidic and Basic Residues Buried at the Same Internal Location in a Protein Are Governed by Different Factors. *J. Mol. Biol.* **2009**, *389*, 34–47. doi:10.1016/j.jmb.2009.03.039
19. Ascone, I.; Castañer, R.; Tarricone, C.; Bolognesi, M.; Stroppolo, M. E.; Desideri, A. Evidence of His61 Imidazolate Bridge Rupture in Reduced Crystalline Cu,Zn Superoxide Dismutase. *Biochem. Biophysical Res. Commun*. **1997**, *241*, 119–121. doi:10.1006/bbrc.1997.7777

20. Friedman, R. (2011). Ions and the Protein Surface Revisited: Extensive Molecular Dynamics Simulations and Analysis of Protein Structures in Alkali-Chloride Solutions. *J. Phys. Chem. B*, **2011**, *115*, 9213–9223. doi:10.1021/ jp112155m
21. Fitch, C. A.; Karp, D. A.; Lee, K. K.; Stites, W. E.; Lattman, E. E.; García-Moreno, E. B. Experimental pKa Values of Buried Residues: Analysis with Continuum Methods and Role of Water Penetration. *Biophysical J.* **2002**, *82*, 3289–3304. doi:10.1016/s0006-3495(02)75670-1
22. Burnett, J. C.; Kellogg, G. E.; Abraham, D. J. Computational Methodology for Estimating Changes in Free Energies of Biomolecular Association upon Mutation. The Importance of Bound Water in Dimer−Tetramer Assembly for β37 Mutant Hemoglobins. *Biochemistry*, **2000**, *39*, 1622–1633. doi:10.1021/bi991724u
23. Burnett, J. C.; Botti, P.; Abraham, D. J.; Kellogg, G. E. Computationally Accessible Method for Estimating Free Energy Changes Resulting from Site-Specific Mutations of Biomolecules: Systematic Model Building and Structural/Hydropathic Analysis of Deoxy and Oxy Hemoglobins. *Proteins* **2001**, *42*, 355–377. doi:10.1002/1097-0134(20010215)42:3<355:aid-prot60>3.0.co;2-f
24. Cozzini, P.; Fornabaio, M.; Marabotti, A.; Abraham, D. J.; Kellogg, G. E.; Mozzarelli, A. Free Energy of Ligand Binding to Protein: Evaluation of the Contribution of Water Molecules by Computational Methods. *Curr. Med. Chem.* **2004**, *11*, 3093–3118. doi:10.2174/0929867043363929
25. Da, C.; Mooberry, S. L.; Gupton, J. T.; Kellogg, G. E. How to Deal with Low-Resolution Target Structures: Using SAR, Ensemble Docking, Hydropathic Analysis, and 3D-

QSAR to Definitively Map the αβ-Tubulin Colchicine Site. *J. Med. Chem*. **2013**, *56*, 7382–7395. doi:10.1021/jm400954h

26. Hashimoto, H.; Nishioka, M.; Fujiwara, S.; Takagi, M.; Imanaka, T.; Inoue, T.; Kai, Y. (2001). Crystal Structure of DNA Polymerase from Hyperthermophilic Archaeon Pyrococcus Kodakaraensis KOD1. *J. Mol. Biol*. **2001**, *306*, 469–477. doi:10.1006/jmbi.2000.4403
27. Lins, L.; Thomas, A.; Brasseur, R. Analysis of Accessible Surface of Residues in Proteins. *Protein Sci.* **2003**, *12*, 1406–1417. doi:10.1110/ps.0304803
28. Eisenmenger, F.; Argos, P.; Abagyan, R. A Method to Configure Protein Side-Chains from the Main-Chain Trace in Homology Modelling. *J. Mol. Biol*. **1993**, *231*, 849–860. doi:10.1006/jmbi.1993.1331
29. Laughton, C. A. Prediction of Protein Side-Chain Conformations from Local Three-Dimensional Homology Relationships. *J. Mol. Biol.* **1994**, *235*, 1088–1097. doi:10.1006/jmbi.1994.1059
30. Krivov, G. G.; Shapovalov, M. V.; Dunbrack, R. L., Jr. Improved Prediction of Protein Side-Chain Conformations with SCWRL4. *Proteins* **2009**, *77*, 778–795. doi:10.1002/prot.22488
31. Alley, E. C.; Khimulya, G.; Biswas, S.; AlQuraishi, M.; Church, G. M. Unified Rational Protein Engineering with Sequence-Based Deep Representation Learning. *Nat. Methods* **2019**, *16*, 1315–1322. doi:10.1038/s41592019-0598-1
32. Senior, A. W.; Evans, R.; Jumper, J.; Kirkpatrick, J.; Sifre, L.; Green, T.; Qin, C.; Žídek, A.; Nelson, A. W. R.; Bridgland, A.; Penedones, H.; Petersen, S.; Simonyen, K.; Crossan, S.; Kohli, P.; Jones, D. T.; Silver, D.; Kavukcuoglu, K.; Hassabis, D. Improved Protein

Structure Prediction Using Potentials from Deep Learning. *Nature* **2020**, *577*, 706–710. doi:10.1038/s41586-019-1923-7

33. Yang, J.; Anishchenko, I.; Park, H.; Peng, Z.; Ovchinnikov, S.; Baker, D. Improved Protein Structure Prediction Using Predicted Interresidue Orientations*. Proc. Natl. Acad. Sci. USA* **2020**, *117*, 1496–1503. doi:10.1073/ pnas.1914677117

Chapter 3: Novel eIF4A1 Inhibitors with Anti-Tumor Activity in Lymphoma†

Introduction

Oncogenic signaling appears to dominate translational output at virtually every stage of cancer propagation for very specific and distinct cellular phenotypes.[1,2] With technological advancements, there is a growing recognition of selectivity in translational regulation mediated by core components of the mRNA biosynthetic apparatus.[3] The regulation of messenger RNA (mRNA) translation in eukaryotic cells is critical for gene expression. It occurs principally at the initiation phase, primarily regulated by eukaryotic initiation factors (eIFs).[4] eIFs are fundamental for mRNA translation and act as the primary targets of numerous oncogenic signaling pathways to modulate gene expression. Thus, anti-tumor agents that strategically target the core components of protein synthesis and related signaling pathways represent novel therapeutic approaches with the potential to overcome resistance due to intra-tumor heterogeneity.

The most tightly regulated step of protein biosynthesis is the initiation of cap-dependent translation in which initiation factors bind to the 5-prime (5′) 7-methylguanosine (m7G) cap of mature mRNA to launch the translation of open reading frames.[5] Cap-dependent translation is driven by the canonical heterotrimeric eIF4F complex, which catalyzes ribosome recruitment to mRNA and is comprised of eIF4G (scaffold protein), eIF4E (cap-binding protein), and eIF4A (ATP-dependent RNA helicase).[6] eIF4A1 unwinds the

† This chapter has been adapted from Kayastha et al. **2022**, *submitted*.[1] Let it be assumed that any indicated supplementary material here is indicated with its name as it appears in Kayastha et al. **2022**.

secondary structure of RNA within the 5'-UTR of mRNA, a critical step necessary for the recruitment of the 43S preinitiation complex, and thus plays a vital role in initiating access to protein biosynthesis for the ribosomes.[7] There are two mammalian isoforms of eIF4A involved in translation: eIF4A1 and eIF4A2.[8] The expression levels of eIF4A1 and eIF4A2 vary in a tissue-dependent manner. eIF4A1 is expressed more in proliferating cells compared to eIF4A2, which is dominantly expressed in growth-arrested differentiated cells, suggesting differential regulation of cell fate.[9] This observation that eIF4A1 and eIF4A2 have distinct biological functions in translational regulation in different subsets of cells and clinical conditions has supported the view that eIF4A1 is a rational cancer target.

The plethora of biochemical data on eIF4A1 (now referred to as eIF4A) reported in the last three decades has deciphered the detailed molecular mechanism of duplex destabilization by eIF4A and the governing principles of its minimal RNA helicase activity.[10] Genome-wide studies of the eIF4A-mediated translatome revealed that helicase regulates the expression of mRNAs encoding vital proteins associated with cell proliferation, cell survival, cell cycle progression and angiogenesis.[11,12] Critically, several reports emphasize that high expression levels of eIF4A significantly stimulate a cancer cell malignant phenotype (proliferation, invasion, migration and EMT) and inhibit apoptosis.[13-16] Thus, the effect of eIF4A up-regulation upon transformed cells appears to act via specific messages, perhaps in addition to a global up-regulation of translation, making eIF4A an attractive target for therapeutic intervention. In addition to the expected findings that eIF4A-dependent mRNAs contained longer 5′-UTRs with a greater degree of secondary structure, both Modelska et al. and Wolfe et al. observed that 5′-UTRs of eIF4A-dependent mRNAs are enriched with G-quadruplex motifs forming potential.[12,13]

Several natural compounds have been characterized that inhibit cap-dependent translation by specifically inhibiting eIF4A activity. These compounds include hippuristanol,[17] pateamine A (PatA), and silvestrol (a rocaglate or "flavagline").[9] Rocaglate analogs are the most studied eIF4A inhibitors in the field.[18] eFT226 (zotatifin), a structure-guided rocaglamide-inspired inhibitor, has entered Phase I clinical trials for solid tumors.[19] Recently, structural elucidation of a rocaglate [RocA]:eIF4A1:polypurine RNA complex revealed that rocaglates operate as interfacial inhibitors and make indispensable interactions with eIF4A1 and two adjacent RNA purine bases.[18] However, all the compounds appear to act non-specifically upon eIF4A1 and eIF4A2. For our purposes, the existence of this structure represents an important starting point for new drug discovery efforts. Knowledge of the interactions made between RocA and the eIF4A1:RNA complex is indispensable for identifying exploitable features of both the ligand and protein that are important for binding both together and can be mimicked by novel inhibitory ligands discovered through virtual screening. In this study, we report the discovery of three novel eIF4A1 inhibitors using this technique, RBF197, RBF203, and RBF208, which potentially bind to the same pockets as RocA but are chemically different. Although these compounds display a lesser degree of potency in their anti-tumor activity than RocA analogs, the observed effective dosage range of the compounds appears to be non-toxic to the transformed cells, and these molecules have scope for further medicinal chemistry design and development using our extensive molecular modeling data.

Methods

Biological assays were performed by Drs. Forum Kayastha and Bandish Kapadia. For the purposes of this thesis, a greater focus is placed on the molecular modeling studies of

this project, but the diligence of Drs. Kayastha and Kapadia was integral to the success of this project. Assays that supported or were related to molecular modeling studies will be discussed, but further information on remaining assays conducted can be found in other communication(s) on this project, including an article recently submitted to *Molecular Medicine*. Examples of these assays include phosphate release luciferase luminescent and phosphate release assays for screening compounds at various concentrations to determine IC_{50}s, and an RNA unwinding assay

Molecular Modeling Studies

Prior to virtual screening, the crystal structure of the eIF4A1 complexed with a polypurine mRNA and RocA (PDB ID: 5ZC9) was obtained from the RCSB Protein Data Bank.[20] RocA pharmacophore-based virtual screening of the MolPort and ZINC15 databases was conducted using the Unity module of the Sybyl-X 2.1.1 suite using its 'Flex Search' option. All docking studies were conducted in GOLD[21] using the built-in ChemPLP scoring function. The high-throughput docking of our virtual screening hits was performed with aromatic ring center constraints and hydrogen bond donor and acceptor constraints, according to their positions in the original virtual screening pharmacophore. Higher-resolution dockings allowed at least 50 solutions per ligand while maintaining all but the original hydrogen bond donor constraints. To validate our docking protocol, RocA was extracted and successfully redocked into the structure of eIF4A1. Post-docking steepest descent energy minimizations were conducted using the Tripos force field in Sybyl-X 2.1.1 with a gradient of 0.02 kcal/mol, 100,000 iterations, Gasteiger-Hückel charges, and a dielectric constant of 8.0. Secondary scoring of eIF4A1-ligand complexes was conducted using the HINT force field,[22,23] a tool developed in our laboratory.

Reagents

The RBF series small molecules were procured from MolPort, Inc platform, silvestrol: Medchem express, WST1: Dojindo Molecular Technologies Inc, phenazine ethosulfate, DMSO: Sigma-Aldrich, D-Luciferin, potassium Salt: Gold Biotechnology. All the other chemicals were procured from Fisher Scientific.

Analysis of eIF4A1 expression in publicly-available DLBCL datasets

UACLAN (http://ualcan.path.uab.edu/) is an extensive resource to evaluate tumor data, primarily The Cancer Genome Atlas (TCGA). The expression of eIF4A1 in TCGA DLBCL samples (n=41) was extracted using the UACLAN database.[24,25] We obtained eIF4A1 in naive B cells from healthy individuals (n = 91) from the DICE [Database of Immune Cell Expression, Expression quantitative trait loci (eQTLs), and Epigenomics] database, which is a comprehensive resource of expression and epigenomic profiles of different types of human immune cells.[26] We also mined the expression of eIF4A1 in other publicly-available DLBCL datasets,[27] GSE10846,[28,29] and GSE87371.[30.31] Finally, prognostic implications of eIF4A1 in DLBCL were assessed using the GDC dataset.[27]

Statistics

Data were analyzed using GraphPad Prism 9. Values were expressed as mean ±S.D. of a minimum of three independent experiments. Wilcoxon signed-rank test was used to compare the data sets between naïve GCB B-cells and DLBCL samples; $p<0.05$ was considered significant. The unpaired Student's *t*-test was used to compare the two groups. One-way ANOVA followed by either Dunnet's or Bonferroni's *post hoc* analysis compared more than two groups; $p <0.05$ was considered significant. Hill coefficient was calculated for the concentration-response curves.

Results

Expression of eIF4A1 Predicts Poor Survival in Diffuse Large B Cell Lymphoma. Before exploring inhibition of eIF4A1, we wanted to first validate it as a potential therapeutic target by assessing eIF4A1's role in lymphomagenesis. To elucidate the pathophysiological relevance of eIF4A1 in DLBCL, we examined publicly available datasets. Analyzing the expression profile of eIF4A1 in the Database of Immune Cell Expression (DICE), Expression quantitative trait loci (eQTLs), Epigenomics,[26] and The Cancer Genome Atlas (TCGA)[24] datasets, we observed a robust increase (p<0.0001) in the transcript levels of eIF4A1 in DLBCL samples compared to naïve B-cells (Figure 3.1A), supporting the relevance of eIF4A1 in lymphomagenesis. Given the substantial variability and unique heterogeneity/biology within DLBCL, this lymphoma subgroup is further classified as Activated B-cell (ABC), Germinal Center B-cell (GCB), and Unclassified (UNC) DLBCL based on its expression profile.[32] In support of our observation, ABC-DLBCL cohorts (n=260) (which have a worse outcome when subjected to standard immune-chemotherapy compared to GCB-DLBCL[33]) display higher expression of eIF4A1 compared to GCB-DLBCL (n=138) (p=0.032) or UNC-DLBCL (n=104) (p=0.191) (Figure 1B). In agreement with these data, transcriptomic profiles (GSE10846[28] and GSE87371[31]) showed that eIF4A1 mRNA was expressed at a higher level in ABC-DLBCL (n=250) compared with GCB-DLBCL (n=268) or UNC-DLBCL (n=64) (Supp. Figure 1).[34] To further validate this observation, we stained primary DLBCL specimens from commercially procured DLBCL tissue microarrays (US Biomax., Inc). In coherence with the above data, the protein levels of eIF4A1 were robustly detected in DLBCL samples (n=377) compared to Reactive Lymph Nodes (LN) (n=54) (Figure 3.2). Staining DLBCL samples

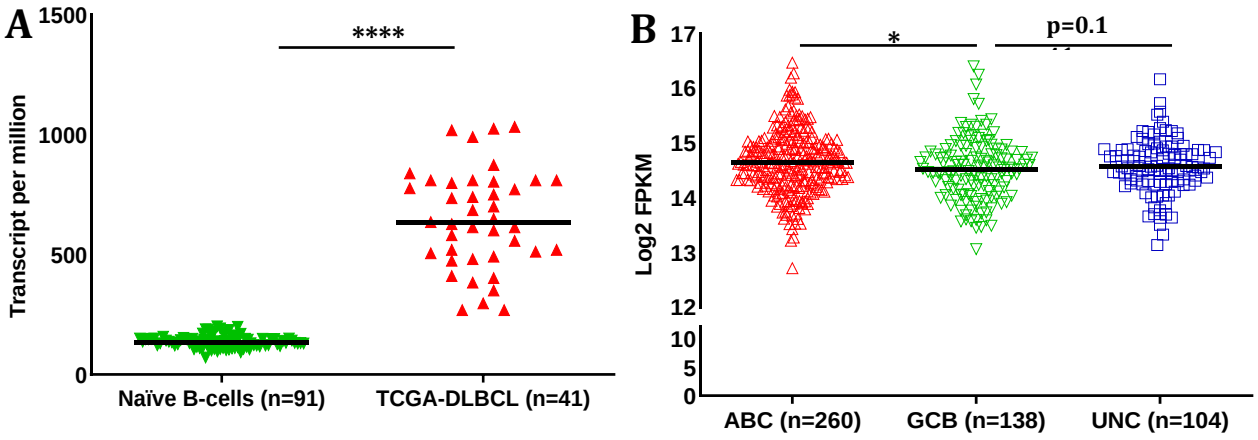


Figure 3.1. Clinicopathologic evaluation of eIF4A1. A) Representative plots show RNA-seq expression profiles of eIF4A1 in naïve B-cells (n=91) (obtained from DICE database https://dice-database.org/) compared with DLBCL (n=41) in TCGA dataset. eIF4A1 showed significantly lower expression in tumor samples compared with control. The Y-axis represents transcript per million (TPM) values. **** p<0.0001 B) Comparison of RNA-seq data of eIF4A1 in molecular subgroups using a publicly available large dataset of patients with DLBCL (https://gdc.cancer.gov/about-data/publications/DLBCL-2018). eIF4A1 showed significantly higher expression in ABC-DLBCL (n=260) subgroups compared with GCB-DLBCL (n=138) and UN-DLBCL (n=104), *p<0.05. The values are represented in log base 2 of fragments per kilobase of exon per million mapped fragments (FPKM).

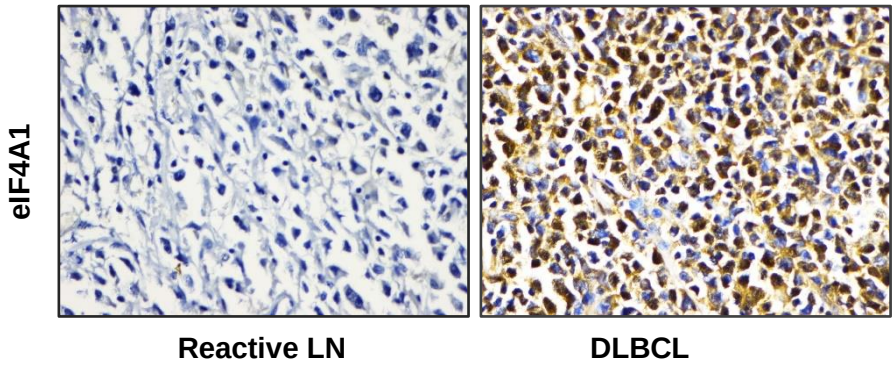


Figure 3.2. Representative immunohistochemistry image of commercially procured (US Biomax., Inc) TMA slides stained with eIF4A1 antibody. Representative scatter plots showing the stained signals of eIF4A1 in reactive lymph nodes compared to DLBCL samples. Statistical analysis was performed using Wilcoxon signed-rank test (unpaired two-tailed), ****$p < 0.001$ vs. reactive LN. Summary chart for DLBCL and normal reactive lymph node samples. -ve: no staining detected, low: 1–2 staining density, high: 3–4 staining density.

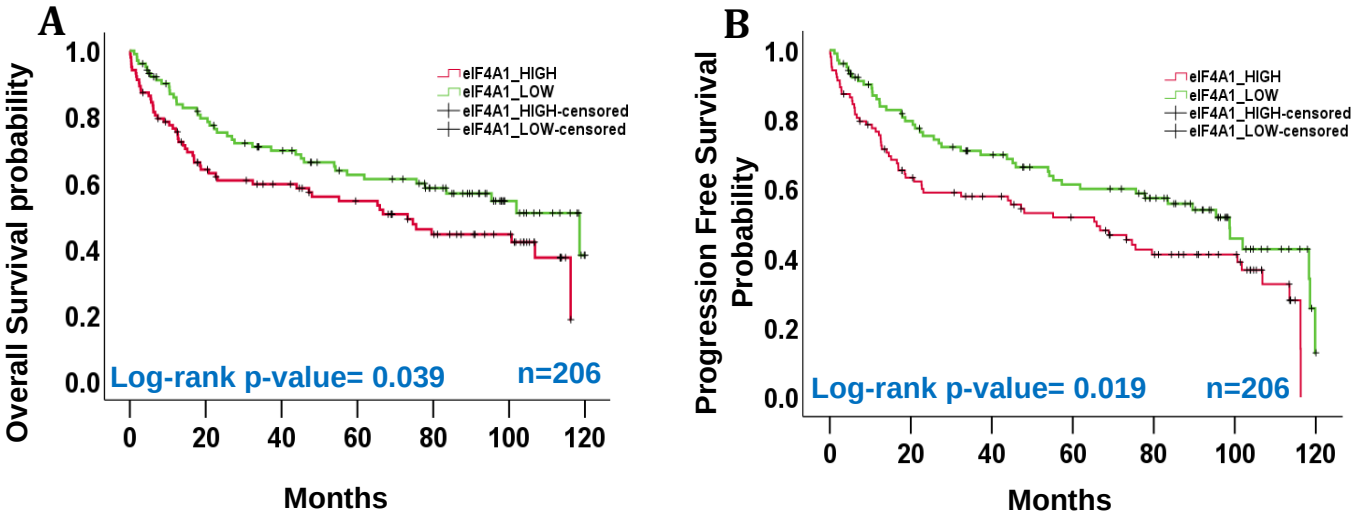


Figure 3.3. Survival rates of patients with expression of eIF4A1. A) eIF4A1 expression was found to be significantly (p=0.039) associated with OS of patients with DLBCL in the publicly available dataset (n=206). Patients with a lower median expression of eIF4A1 showed a better prognosis than patients having higher median expression. B) eIF4A1 expression was also found to be significantly (p=0.019) associated with the PFS in the same cohort of patients with DLBCL having a similar observation.

displayed 72% expression of eIF4A1 while that of reactive lymph node was 33%. Collectively, eIF4A appears to be upregulated in DLBCL.

To further investigate the clinical importance of eIF4A1 in lymphoma progression, the prognostic value of eIF4A1 gene expression was determined using publicly available datasets,[24-26] employing a cox p-value<0.05. As shown in Figures 3.3D and 3.3E, eIF4A1 was significantly associated with overall survival (OS) and progression-free survival (PFS) of patients with DLBCL. Patients with a higher median expression of eIF4A1 showed shorter survival periods than those with lower median expression. (Figure 3.3A, n=206, p=0.039). Consistently, patients with higher expression of eIF4A1 have a shorter progression-free interval than patients with low expression of eIF4A1 (Figure 3.3B, n=206 p=0.019). Altogether, the above clinical data endorses the claim that higher eIF4A1 gene expression is

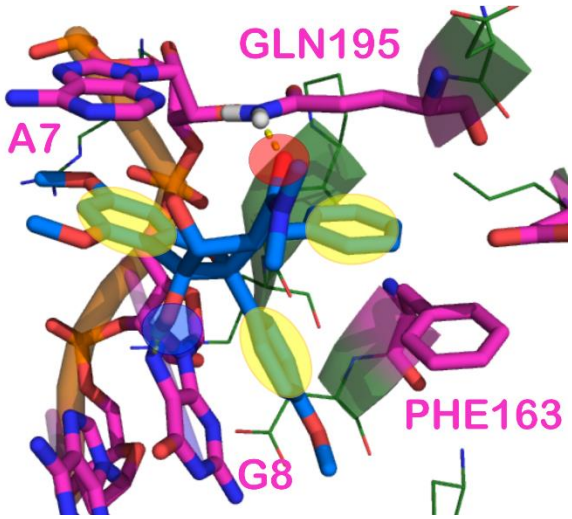


Figure 3.4. Model of RocA used to define important pharmacophore features used in pharmacophore-based virtual screening experiment. In yellow circles are shown regions defining positioning of aromatic rings. The red circle indicates a hydrogen bond acceptor interacting with GLN195, while the blue circle indicates a hydrogen bond donor interacting with G8.

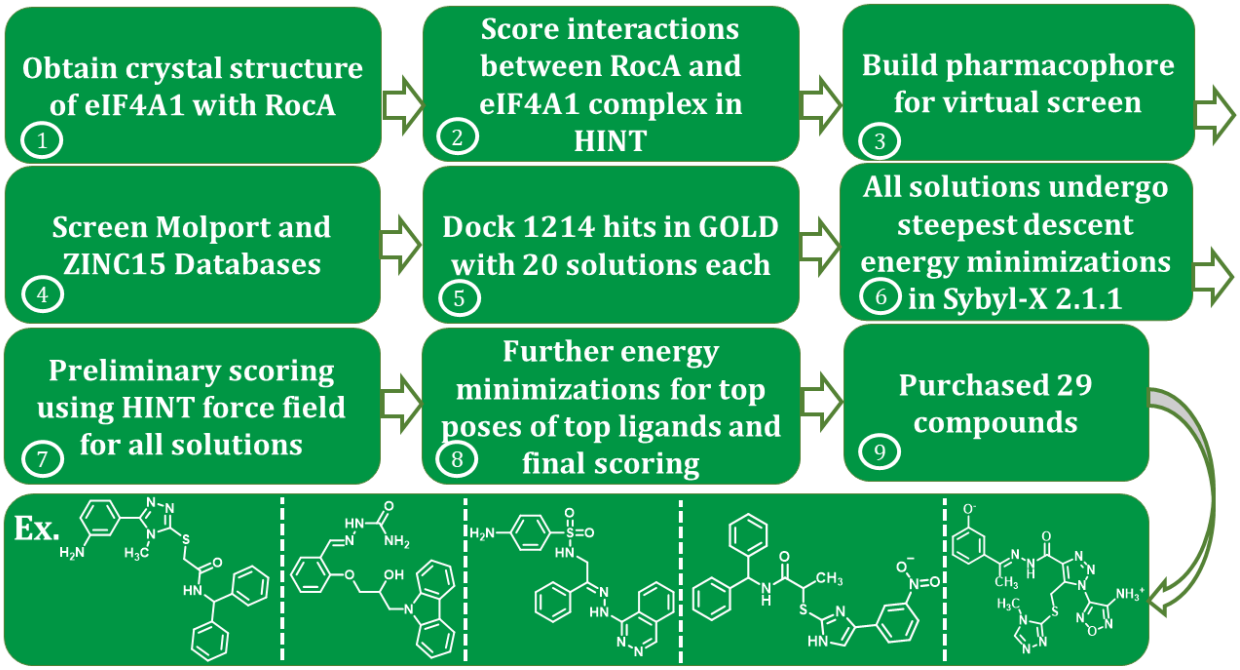


Figure 3.5. Workflow for virtual screening strategy that identified RBF98 as the top hit. Stages for this workflow included obtaining the crystal structure of eIF4A1 complexed with RocA, scoring interactions between these two species, constructing and implementing the virtual screening pharmacophore, high-throughput molecular docking, energy minimizations of solutions, preliminary scoring of solutions in HINT, and final energy minimizations and scoring, followed by the purchase of the 29 top-scoring hits.

associated with poor survival and more aggressive clinicopathological features, supporting our notion that eIF4A1 is a promising therapeutic target in DLBCL.

Structure-Based Drug Screen Identifies New Inhibitors of eIF4A1

Our search for novel inhibitors of eIF4A1 began with a structure-based virtual screen of a potential binding site of eIF4A1. Using a crystal structure (PDB ID: 5ZC9[35]) of eIF4A1 complexed with a polypurine mRNA strand and RocA, which is a natural product inhibitor of eIF4A. To establish what features of RocA to use for the basis of our virtual screen, we utilized the HINT force field[22,23] to determine the molecular features of RocA most responsible for its binding. Briefly, HINT is a scoring function based on the free energy associated with solvent partitioning between 1-octanol and water (see Chapter 1). It has been used in numerous studies involving interactions between and amongst proteins, polynucleotides, and small molecules.[36-38] In addition to π-π stacking interactions between the mRNA bases of the polypurine strand and with PHE163 in literature reported by Iwasaki et al.,[35] we noted two

Table 3.1. Major interactions identified between RocA and eIF4A1:RNA complex with HINT

Residue/nucleotide	Residue/nucleotide Atom	RocA Atom	Interaction Score	Interaction Type
GLN195	NE2	O1	240	Hydrogen Bond
G8	N7	O6	87	Hydrogen Bond

key hydrogen bonding interactions involving the ligand and GLN195, as well as G8 (Figure 3.4). Table 3.1 lists these crucial interactions that we concluded contributed most to RocA's binding and activity.

The hydrogen bond between O6 of RocA and G8 of the mRNA is consistent with Iwasaki et al.'s structural studies. Thus, the aromatic rings of RocA, along with its acceptor

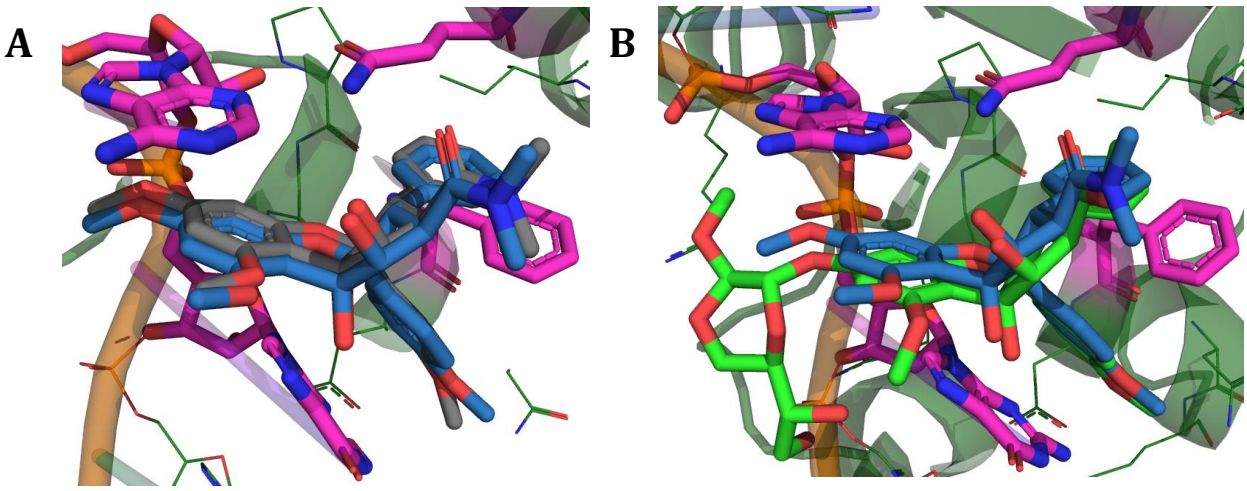


Figure 3.6. Top-scoring docked poses in HINT and GOLD for RocA and Silvestrol, respectively. Both poses were obtained with the same docking protocol to validate the method used to dock hits obtained from our virtual screening. A) The docked pose of RocA (in gray) overlaps almost exactly with the co-crystallized structure of RocA (in blue). B) Features shared between Silvestrol (in lime green) and RocA (in blue) overlap extremely well.

O1 and donor O6, were used as pharmacophore features in a virtual screen for new inhibitors (Figure 3.4). Our overall virtual screening protocol is summarized in Figure 3.5. Virtual screening was performed using the 'Flex Search' option of the Unity[39] suite in Sybyl-X 2.1.1. A total of 1218 hits from the screen underwent high-throughput docking in GOLD 5.6.1 with 20 solutions per ligand. In order to validate our docking approach, RocA was extracted and successfully redocked into the structure of eIF4A1 where it overlapped well with its co-crystallized pose (Figure 3.6A). All solutions were triaged into the HINT force field for secondary scoring. The top 29 scoring compounds (Supp. Table 2)[34] from HINT were purchased and further assayed for activity. To rapidly evaluate the inhibitory ability of the selected novel eIF4A inhibitors, we took advantage of an in-cell high throughput eIF4A-3X luciferase assay.[12] The luciferase-based reporter assay with 5'UTR of eIF4A-sensitive four tandem repeats of the $(CGG)_4$ 12-mer motif (GQs) driven by beta-actin promoter was used as a platform for primary screening in Hek293T/17 stables cell lines (Figure 3.7). Cofactors like eIF4B stimulate the activity of eIF4A1. However, the eIF4A regulated luciferase readout

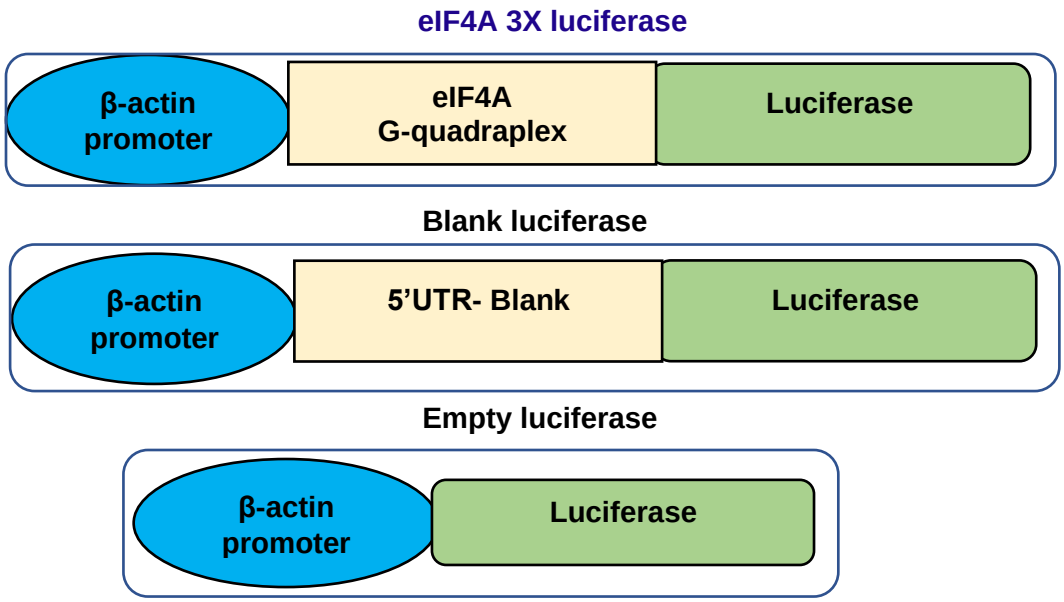


Figure 3.7. Design of luciferase construct with 5'UTR of eIF4A1 G- quadruplex sequence with the β-actin promoter, negative controls, blank with scrambled sequence and empty test construct.

was minimally dependent on eIF4B (Supp. Figure 2B).[34] Similar experiments were performed with silvestrol as an internal control (Supp. Figure 2C),[34] suggesting that the consensus sequence is highly reliant on eIF4A. Silvestrol, structurally, shares a scaffold similar to RocA. To verify that it likely binds in a manner similar to RocA, silvestrol was docked into the structure of eIF4A1, where its highest scoring pose overlapped well with the co-crystallized pose of RocA (Figure 3.6B). The 29 commercially available hit compounds were used for the initial primary screen at concentrations ranging from 1 nM to 10 μM. Luciferase readout greater than 50% was considered the cut-off value for the screen. We observed that RBF98 (showed around 50% inhibition with respect to the DMSO control at 1 nM concentration (Figures 3.8A-3.8C). Interestingly, the percentage decrease in the luciferase readout was less than 10% in blank and empty luciferase groups, suggesting high specificity for the compound (Supp. Figure 3A)[34] in limiting eIF4A-driven translation. It should be noted that the compound RBF98, at higher concentrations, showed a decrease in

eIF4A inhibition, probably due to various physicochemical properties such as reduced solubility, etc. The ability of compound RBF98 to preferentially target eIF4A-sensitive luciferase with a readout similar to silvestrol provides promising evidence that RBF98 inhibits eIF4A and not another protein in the general translational apparatus and also binds in a manner similar to silvestrol and RocA. After analyzing its highest-scoring docked pose, it was concluded that RBF98 might adopt a similar binding mode to RocA, with three of its aromatic rings forming π-π stacking interactions with two adjacent nucleotide bases and PHE163, which seem to be crucial for rocaglate activity. Further, this docked pose of RBF98 shows that its phenoxide moiety may form a hydrogen bond with G8 of the RNA strand and a novel ionic interaction between its ammonium group and ASP198. This previously unobserved interaction may be integral for achieving improved drug-like properties over rocaglamate-based inhibitors (Figure 3.9). Additional biochemical testing was performed with RBF98 to investigate the direct inhibition of the compound on eIF4A's helicase activity. Here, we ran an inorganic phosphate release assay to directly measure the eIF4A1 ATP-dependent RNA helicase activity using a stable mixture of yeast RNA. To achieve a maximal signal-to-noise ratio in this endpoint assay, we optimized the amount of enzyme and the incubation time (Supp. Figures 3B, 3C).[34] RBF98 showed a dose-dependent decrease with an inhibitory effect at ~50% at a concentration of 0.3 μM compared with the DMSO control (Figure 3.10).

We next analyzed RBF98 impact on a cellular proliferation assay. The compound reduced the cellular proliferation of DLBCL cells at 0.5 μM and 1 μM concentrations. Silvestrol was again used as a positive control (Figure 3.11). Cellular proliferation at lower

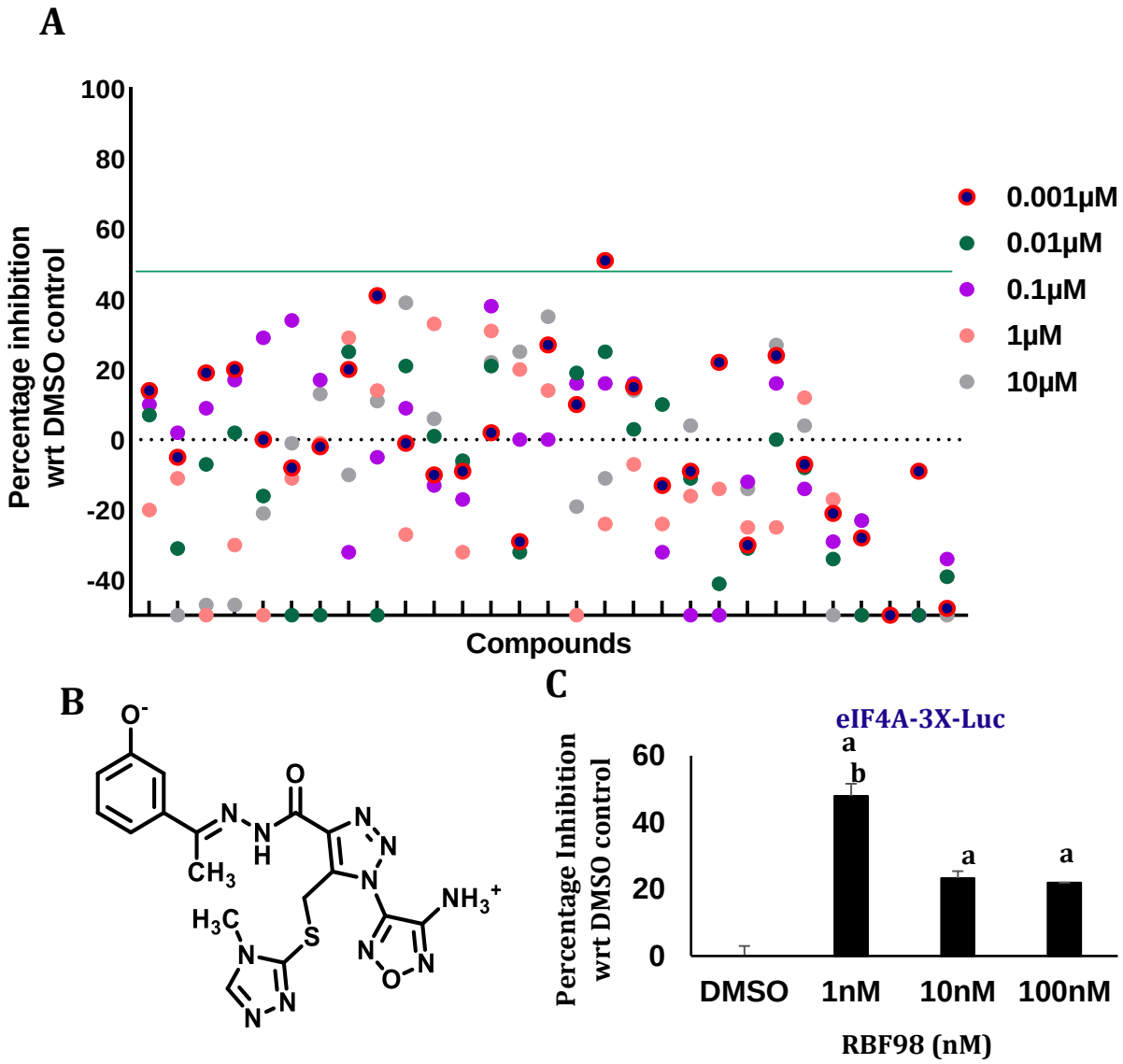


Figure 3.8. eIF4A1 specific high throughput screen identifies small molecules with inhibitory effect. A) Scatterplot of primary screen results. A total of 29 compounds were tested and luciferase signal reduced by ≥50% compared to control were identified and considered active. Luciferase activity results are expressed relative to values obtained in the presence of vehicle controls. Percentage inhibition was calculated and plotted in a scatter plot, n=3 biological replicates performed ±SEM. B) Structure of RBF98, a candidate inhibitor. C) Percentage inhibition was observed in the treatment of RBF98 at various concentrations in eIF4A1-3X-Luciferase Hek293T/17. Treatment groups vs DMSO control groups $^{a}p< 0.05$; $^{c}p<0.001$, $^{d}p< 0.0001$. Experimental groups vs 1mM treatment groups $^{\alpha}p<0.05$. $^{\beta}p<0.001$. $^{¥}p<0.0001$.

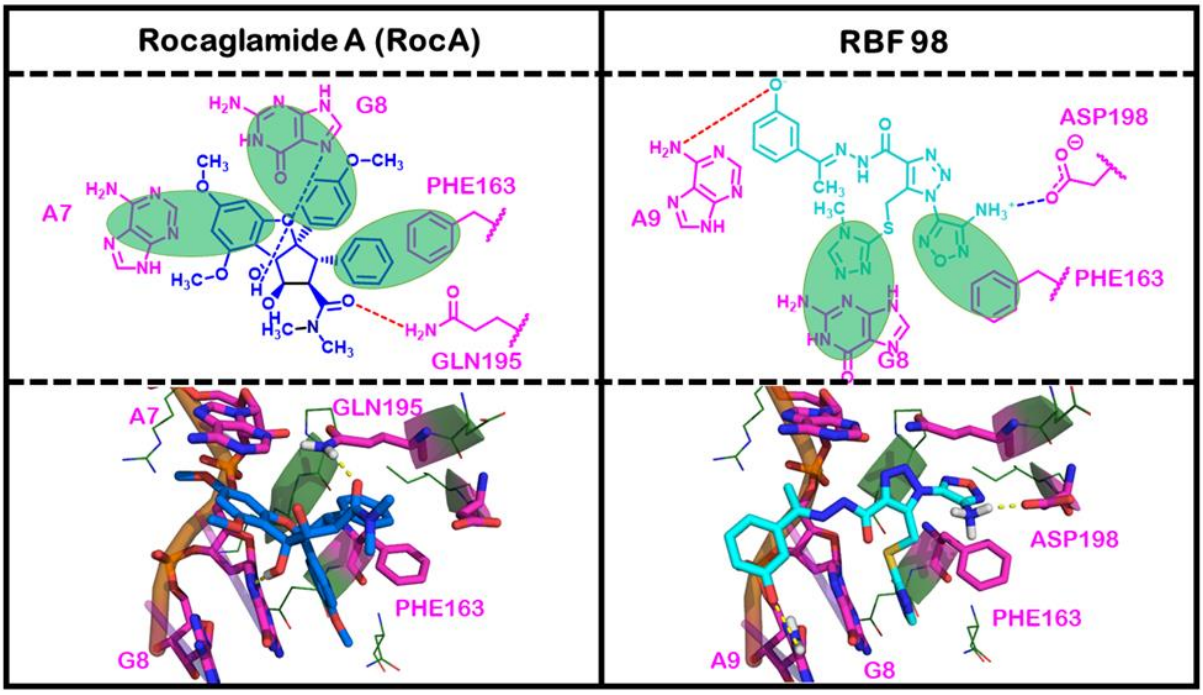


Figure 3.9. Interaction environments for RocA and RBF98. The above two panels show stick representations of the interactions made between RocA and RBF98 and their surrounding environments. Green, transparent ovals are used to two-dimensionally represent possible π-π stacking interactions between the ligands and surrounding residues. RocA forms these π-π stacking interactions with the A7 and G8 bases and PHE163, while RBF98 forms them with only G8 and PHE163. Dashed lines between the ligands and surrounding residues are used to indicate hydrogen bonding, where the color indicates the donor/acceptor character of the ligand atom (blue = donor; red = acceptor). RocA donates a hydrogen bond to G8 and accepts on from GLN195, while our model of RBF98 accepts a hydrogen bond from A9 and donates an ionic interaction to ASP198, a potentially unobserved interaction.

concentrations had minimal impact on DLBCLs (data not shown). For further insight into the effect of RBF98, we performed a colony formation assay in the OCI-Ly3, Toledo (malignant cell lines), and lymphoblastoid cells (GMO 17220B, 1528, 13604, non-malignant cell lines). We observed that RBF98 significantly decreases colony formation in malignant cell lines in a dose-dependent manner (Figures 3.12A, 3.12B, Supp. Figure 4B).[34] Notably, treatment of lymphoblastoid cells had minimal impact on their proliferative capacity indicating that the compound may have a potential non-toxic effect on non-malignant cells, addressing a major limiting factor of the currently available eIF4A inhibitors (Fig. 3.12A,

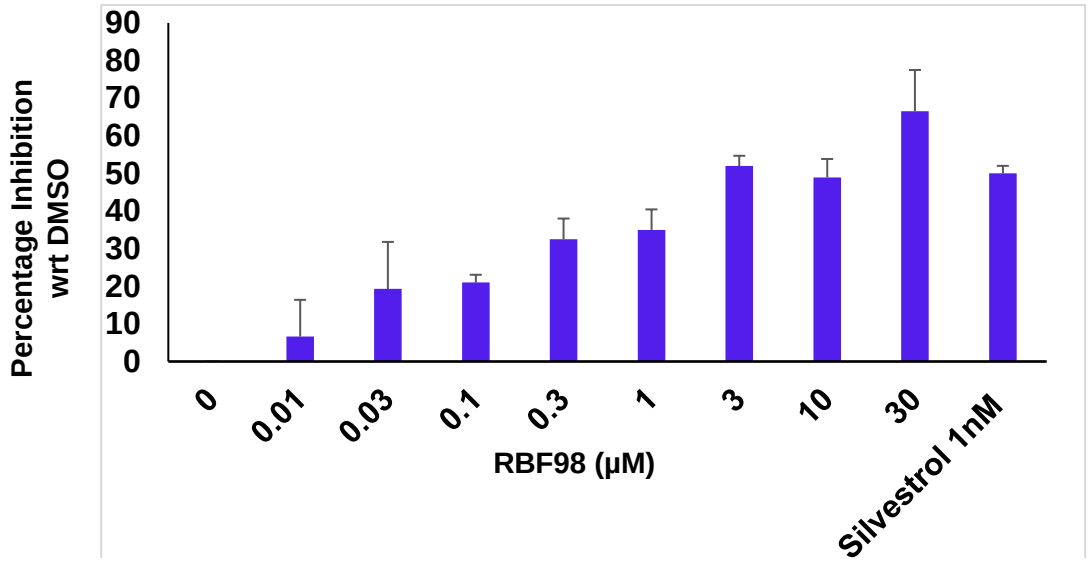


Figure 3.10. Dose-dependent percentage inhibition of human eIF4A1 *in-vitro* activity on the treatment of RBF98, compared to DMSO control using an inorganic phosphate release assay (SensoLyte kit). IC_{50} values observed were observed to be 3 μM.

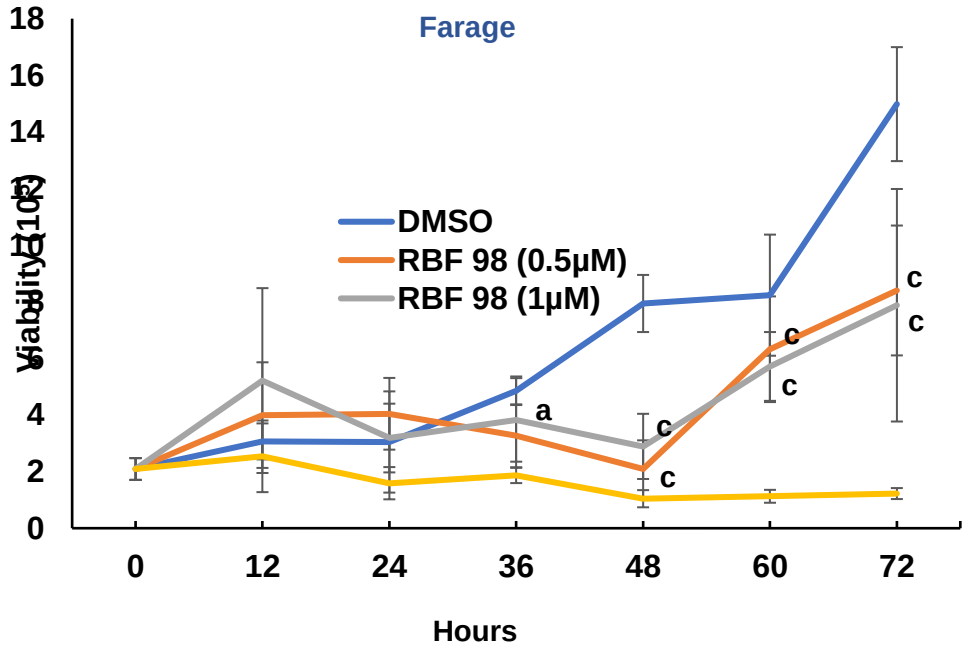


Figure 3.11. For our proliferation assay, Farage (GCB) origin was seeded at a density of 10,000 and treated with 0.5 and 1μM of RBF98 for up to 72 hours. The cell viability was measured at different time points using the trypan blue method. Silvestrol treatment was done at 50nM as a positive control group. Viability was observed to be decreasing with increasing time in comparison to DMSO control ($^{a}p< 0.05$; $^{c}p<0.001$, $^{d}p< 0.0001$)).

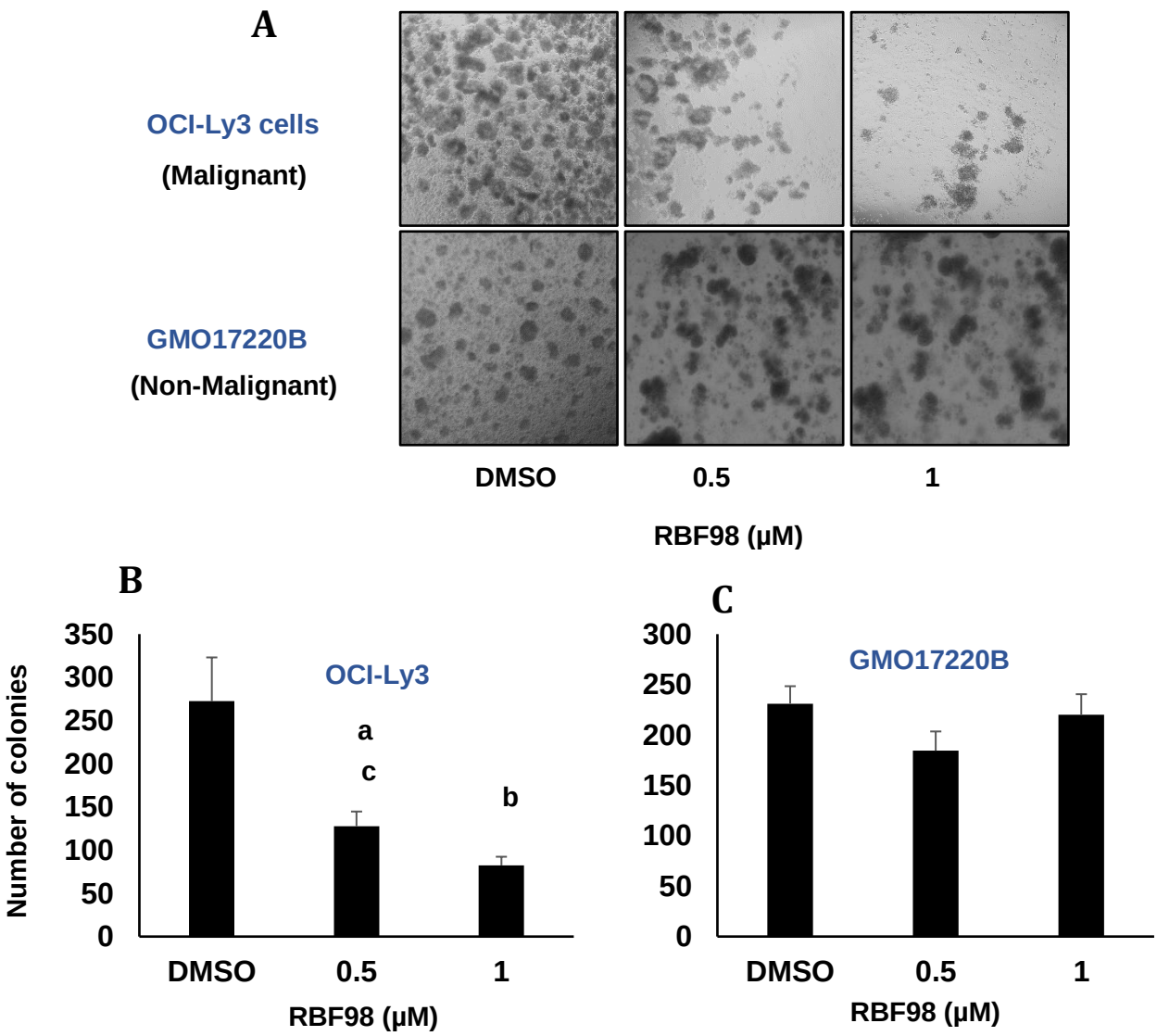


Figure 3.12. Effect of RBF98 on DLBCL colony formation. A) Representative image of the colony formation in OCI-Ly3 (malignant) and GMO17220B (non-malignant) cells. The total number of colonies grown in B) OCI-Ly3 and C) GMO17220B cells upon treatment with 0.5 and 1 μM of RBF98. Statistical analysis was performed using one-way ANOVA followed by Bonferroni correction analysis. For p values, see Figure 3.8.

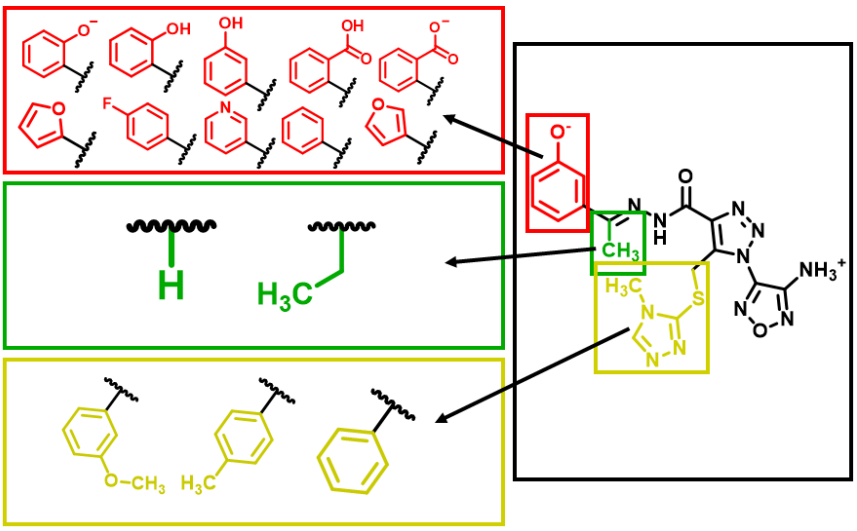


Figure 3.13. Summary of compounds sampled in secondary screen. Boxes surrounding different moieties of RBF98 correspond by color to the larger boxes containing functional groups that were sampled as part of this secondary screen in different combinations. Purchased analogues were selected based on Tanimoto index similarity to RBF98. In black boxes are the three top hits resulting from this screen: RBF197, RBF203, and RBF208 with their IC_{50} values in our luciferase assay.

3.12C, Supp Fig. 4B).[34] To further explore the molecular insights of RBF98's activity, we pulse-labeled DLBCL cells with puromycin after treatment with the compound. Immunoblotting with anti-puromycin revealed a concentration-dependent decrease of puromycin labeling along with the protein levels of eIF4A, but minimal changes in eIF4E, indicating an overall reduction in the translation capacity of the cells (Supp Fig. 4C).[34] Further, the expression of eIF4A-dependent genes, cMYC[40] and CyclinD1,[41] was reduced similarly (Supp Fig. 4C).[34] After screening the initial set of 29 compounds, we pursued analogs of RBF98 in the hope of identifying purchasable compounds with better or comparable activity to it, our most potent hit. Using a similarity search function based on Tanimoto indices[42] built into the MolPort website, we identified 34 analogs (Supp. Table 3)[34] that had chemical similarities to RBF98. Generally, the compounds from this round of screening structurally differed in three positions from RBF98, which allowed us to probe

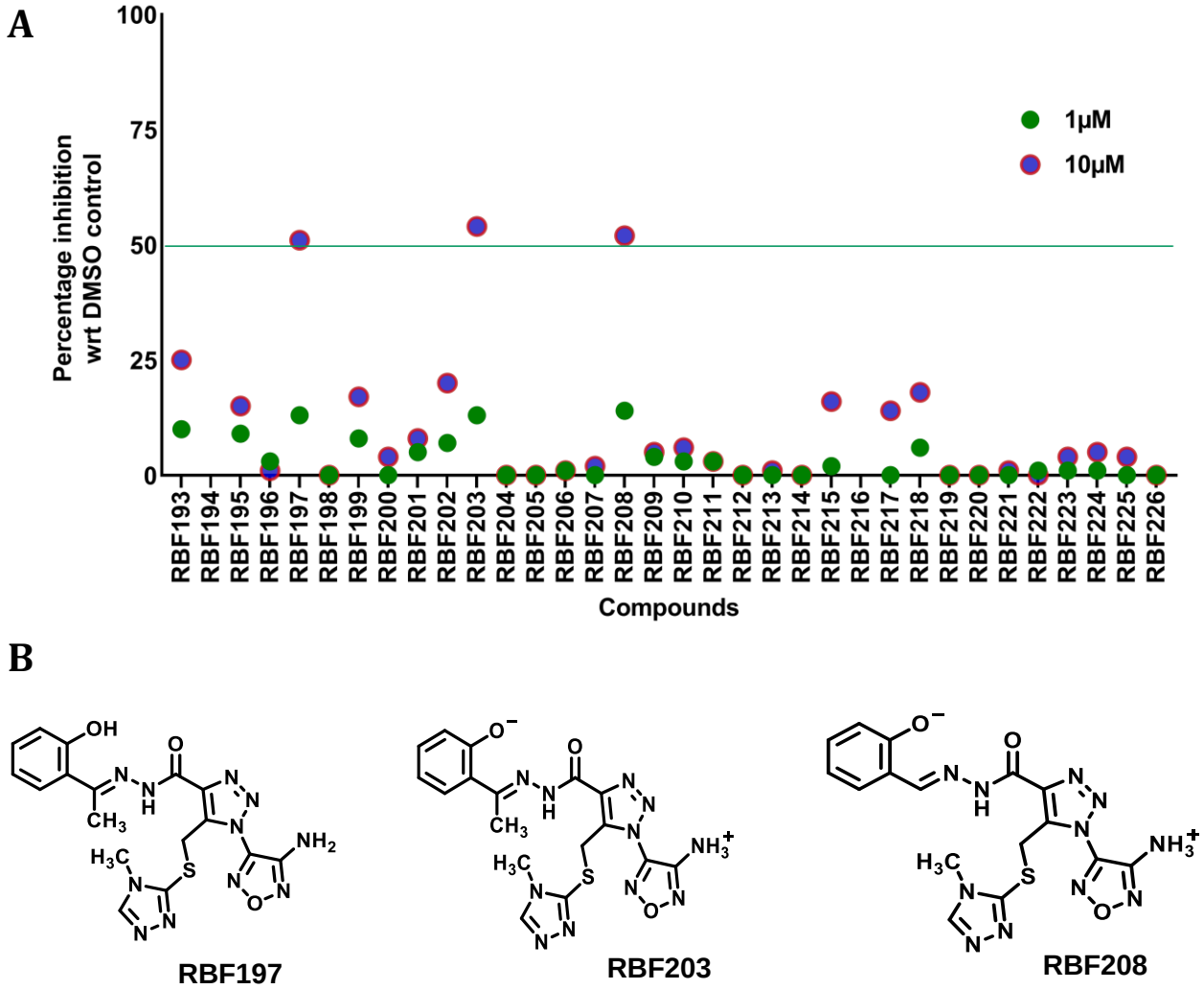


Figure 3.14: Secondary screen of RBF98 analogs in eIF4A1-3X-luciferase Hek293T/17. A) A total of 34 compounds that inhibited Luciferase signal by ≥50% compared to control were identified. Luciferase activity results are expressed relative to values obtained in the presence of vehicle controls. Percentage inhibition was calculated and plotted in a scatter plot, n=3 biological replicates performed ±SEM. B) Structures of RBF197, RBF203, and RBF208, potent candidate inhibitors.

aspects of the structure-activity relationship of our hit (Figure 3.8A, 3.8B). We mainly focused on altering these positions because they were the most accessible changes, based on the collection of compounds commercially available from MolPort. Figure 3.13 illustrates some of the different structural variations in these positions. After the luciferase readout, the primary screen was applied to the 34 compounds we obtained, including RBF197, RBF203,

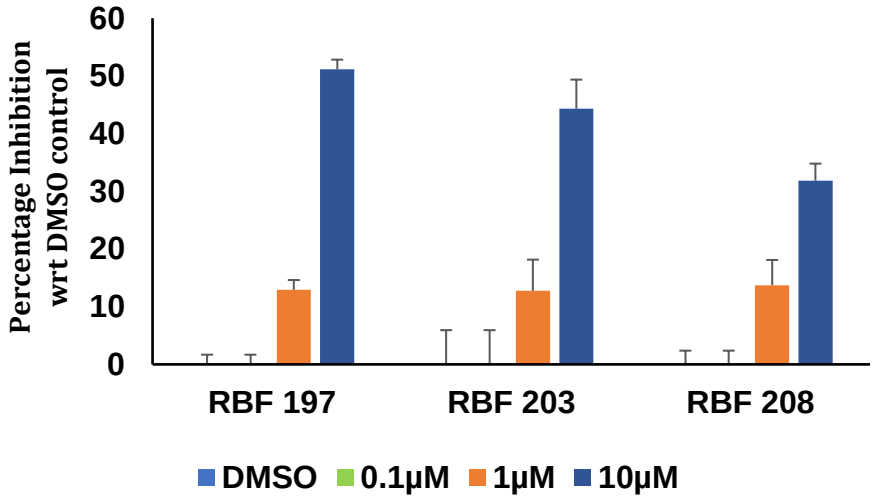
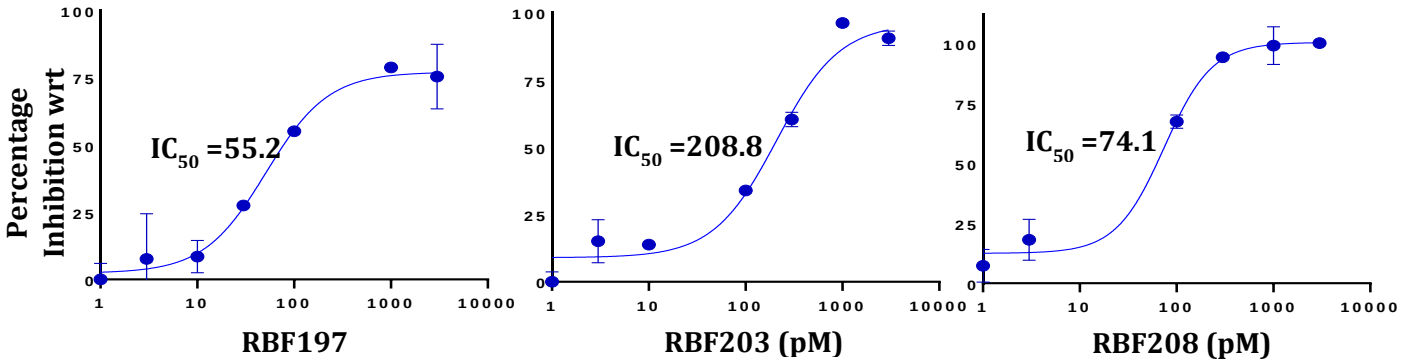


Figure 3.15. Representative plots of percentage inhibition values of luciferase activity on the treatment of RBF 197, 203, and 208 at 0.1, 1, and 10 μM in eIF4A1-3X-Luciferase in Hek293T/17 cell lines for 24 h (n=3).

A



B

Hill Coefficient	
RBF 197	1.35
RBF 203	1.33
RBF 208	1.75

Figure 3.16. Percentage inhibition of human eIF4A1 *in-vitro* activity on the treatment of RBF197, RBF203, and RBF208 in inorganic phosphate release assay (SensoLyte Kit). A) Concentration-response curves of RBF197, RBF203, and RBF208, compared to DMSO control. IC_{50} values observed were 55.2, 208.8, and 74.1 pM, respectively. B) Hill coefficient values for the concentration-response curves.

and RBF208 (Figures 3.14A, 3.14B), which displayed a dosage-dependent decrease in luciferase readout (Figure 3.15). More importantly, all three hit molecules do not show more than 15% inhibition of blank (Supp. Figure 5B)[34] and empty luciferase readout (Supp. Figure 5A),[34] implicating a specific inhibitory effect on eIF4A dependent activity. Most of the compounds available for purchase from our virtual screen and that we assayed showed alterations to the *m*-phenoxide moiety of RBF98, thus allowing us to probe this region extensively (Figure 3.13). The other two regions of the RBF98 first-round lead remain largely unexplored. To further corroborate our findings, we subjected these three new hits to a kinetic assessment using an *in-vitro* inorganic phosphate assay. To our surprise, the selected three compounds potently inhibited eIF4A helicase activity in a dose-dependent manner with IC_{50} values in the picomolar range and with Hill coefficients 1.35, 1.33 and 1.75 indicating single-molecule binding without aggregating effects (Figures 3.16A, 3.16B).

Novel eIF4A Inhibitor Blocks Cell Proliferation and Impedes Overall Translation in DLBCL.

To determine if the most potent compounds exert inhibitory activity in DLBCLs, we first subjected a panel of DLBCL cells to a WST-1 cell viability assay. This assay quantitates the number of living and metabolically active cells by measuring the cleavage of tetrazolium salts by intracellular enzymes. As shown in Table 3.2, all hit compounds decreased the cellular viability of DLBCL cells with an EC_{50} in the low micromolar range (Supp. Figure 6A).[34] To our surprise, SUDHL2, which harbors a mutation in A20, SOCS1, and TP53, was insensitive to the eIF4A inhibitors.[43] Additional analogs with varying potencies in the luciferase readout assays were tested in the cell viability assay to investigate if helicase inhibition tracked

Table 3.2. IC_{50} values of RBF197 and RBF208 in a panel of five DLBCL cell lines were performed using WST-1 assay.

DLBCL Cell Line	RBF 197 (μM)	RBF208 (μM)
OCILy3	0.4	0.9
SUDHL4	3.2	4.5
Farage	2.4	2.9
DS	2.8	4.3
RC	1.9	5.2
SUDHL2	>30	>30

DLBCL WST1 inhibition, and to ensure that cell viability did not decrease due to the general toxicity of the inhibitor scaffold (Supp. Table 4).[34] As anticipated, these molecules do not dramatically affect the DLBCL cellular viability; thus, minor substitution to our selected compounds that hampers their eIF4A inhibitory capacity also displays a minor reduction in potency of DLBCL cellular viability.

Given that eIF4A inhibitors demonstrated a statistically significant reduction in cell proliferation in DLBCL subtypes as well as depletion of critical oncogenes, we next performed a colony formation assay with a panel of DLBCL and lymphoblastoid cells. We selected seven different cell lines; five are of DLBCL (ABC[2] or GCB[3]) origin, while the other two are non-malignant cell lines (lymphoblastoid, GMO). In agreement with our previous data, we observed a significant dose-dependent decrease in the number of colonies formed. Notably, RBF197 displayed minimal effect on colony formation in GMO cell types, while

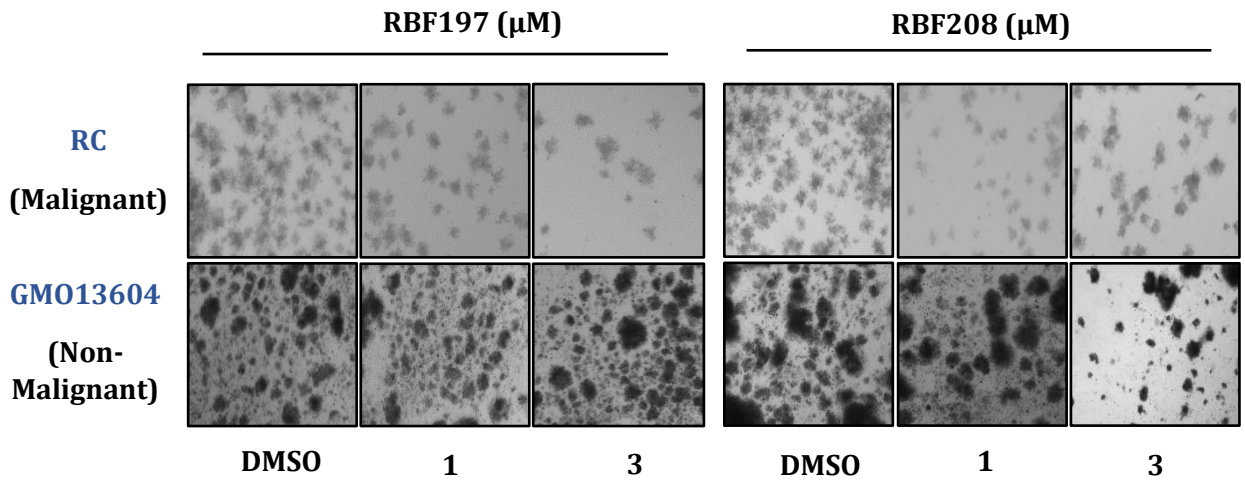


Figure 3.17. Effect of RBF197 and RBF208 on DLBCL colony formation. Representative image of the colony formation in RC (malignant) and GMO13604 (non-malignant) cells.

RBF208 reduced colony formation in GMO cell lines at a higher concentration (Figure 3.17, Supp. Figures 8A, 8B).[34] These results are consistent with RBF197 and RBF208 being selective eIF4A inhibitors, while the therapeutic window for RBF197 is broader than RBF208.

Potential RNA Clamp Mechanism of eIF4A Inhibition

To account for the increased activities seen for RBF197 and RBF208, we performed more extensive docking studies for these compounds. Not surprisingly, considering their flexibility and the size/shape of the pocket, we obtained a wide variety of high-scoring docked poses for these compounds. Logically, we thought the best approach to our modeling may be to identify the accurate binding mode for our most potent compound, RBF197, and use its most probable and favorable pose to deduce the same kind of pose for our less potent compounds and the structural reason(s) for their decreased activity. We attempted higher resolution dockings for RBF197 to this end and progressively refined our docking protocol based on different iterations of dockings. In our first attempt, we sought to determine

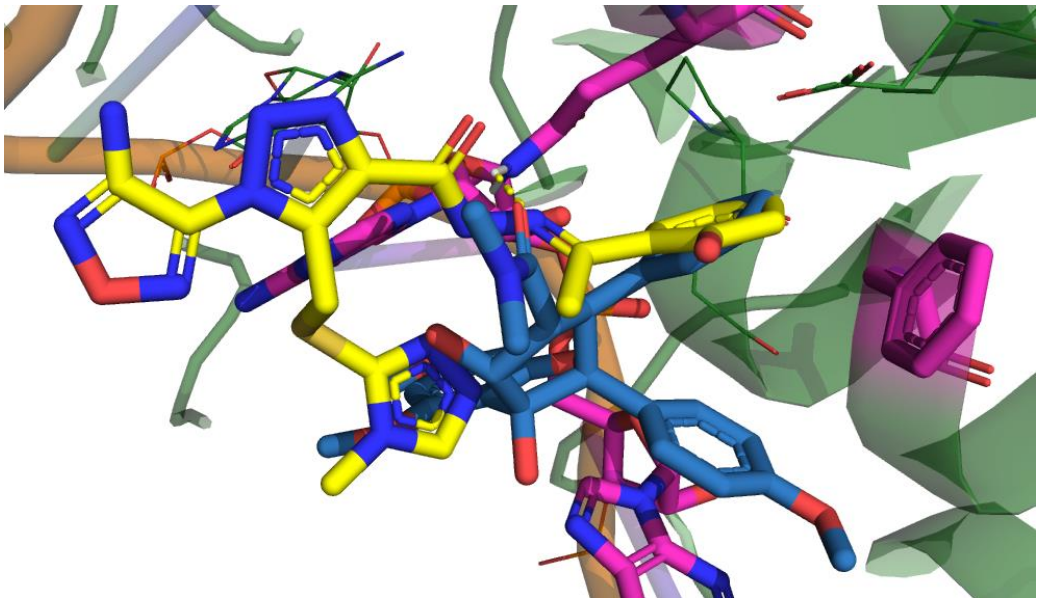


Figure 3.18. Docked pose of RBF197 having used an aromatic ring center and hydrogen bond acceptor constraints of the original virtual screening pharmacophore. This docked pose, although it overlaps well with the constraints used, occupies a very different overall position within the RocA binding site than RocA and does not make the same π-π stacking interactions with A7 and G8 as RocA.

whether all of the VS pharmacophore-based docking constraints were necessary and thus removed all but an aromatic ring feature to interact with PHE163 and a hydrogen bond acceptor feature to interact with the donor end of GLN195's side chain. The best result from this docking is seen in Figure 3.18. Here, RBF197 makes very different interactions from a different position within the RocA binding site than RocA, namely with a nearby phosphate from the RNA chain and now orients its phenol to π-π stack with PHE163. Although this pose scores highly with HINT (Score = 968), this pose makes little use of π-π stacking interactions that are essential for RocA binding and does not effectively utilize its phenol hydroxyl that appears to be responsible for its higher potency compared to its predecessor. We determined, based on this result, that this is not the pose in which RBF197 binds, and all three original aromatic features that were also designed to interact with A7 and G8 are necessary to properly position RBF197 in the RocA binding site before it forms interactions

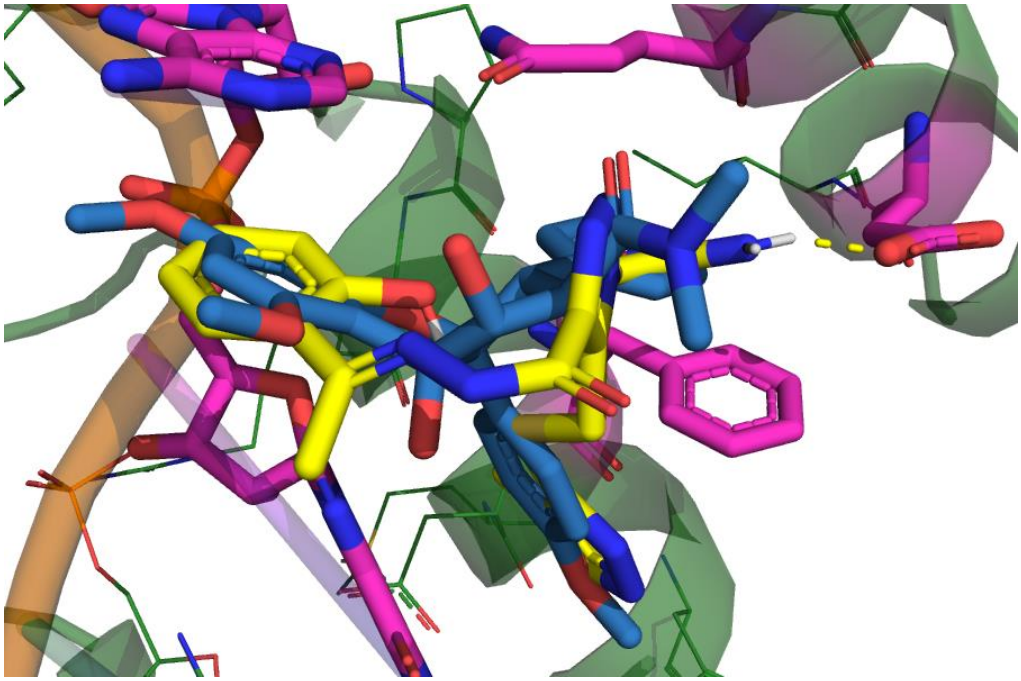


Figure 3.19. Docked pose of RBF197 using three aromatic ring constraints and a hydrogen bond donor constraint used to interact with the acceptor end of GLN195's side chain. This pose has improved π-π stacking compared to the pose in Figure 3.18 and similar to RocA. It retains the hydrogen bonding interaction with ASP198 and offers potential for hydrogen bonding between the *o*-OH of RBF197's phenol and the carbonyl of GLN195's amide.

with nearby residues. In another docking attempt, we restored the original aromatic ring center constraints and received what we believe was a more properly position docked pose of this ligand with better overlap of RBF197's aromatic rings and those of RocA (Figure 3.19). This pose resulted from rotating the amide of GLN195's side chain to experiment with forming a potentially new interaction with its acceptor. Although no interaction was made with it in this pose, the *o*-OH of RBF197's phenol was positioned where it favorably could make such an interaction. Given its better overlap with the aromatic ring pharmacophore features and potential utility for the *o*-OH, we believed this pose to be closer to the real binding mode of RBF197. Our final iteration of docking, after obtaining 300 solutions, resulted in the pose visible in Figure 3.20. This specific pose, indeed, formed a new interaction between the *o*-OH and the acceptor of GLN195's side chain and retained the

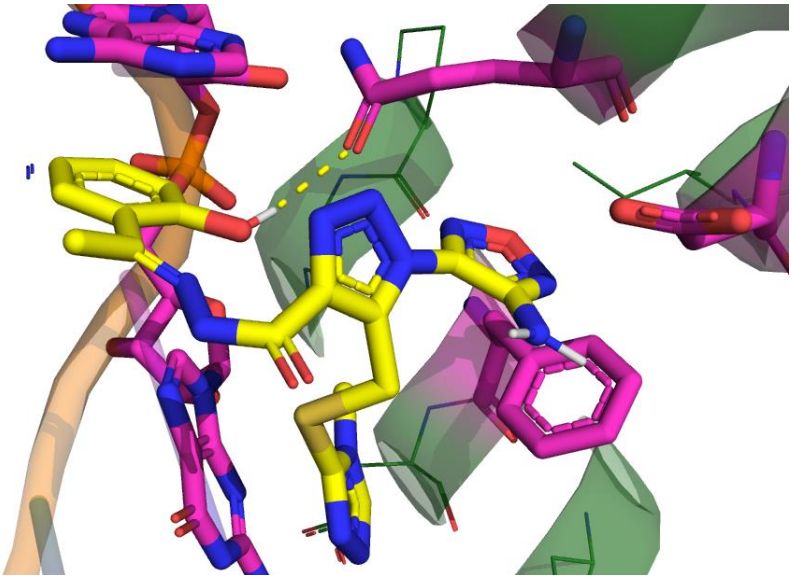


Figure 3.20. Docked pose of RBF197 forming a new hydrogen bond with GLN195 using its *o*-OH. This pose retains previously seen π-π stacking interactions with PHE163 and nearby nucleotides, but not hydrogen bonding with ASP198.

traditional π-π stacking with PHE163 and nearby nucleotides. It did not, however retain the hydrogen bond with ASP198, which we determined was still not impossible by means of manual rotations of certain bonds. Our logic was that manually rotated positions of residues and RBF197 in the binding site, after undergoing energy minimizations, should still be energetically favorable, should they ultimately retain their poses. Residues, such as ASP198, are immersed in solvent and should also be somewhat flexible. Following rotation of ASP198 into positioning for a hydrogen bond with RBF197's amine and energy minimization, positions of interacting species were remained static, resulting in the pose visible in Figure 3.21. This pose was also used as a template for docking RBF208, also visible in Figure 3.18. It should be noted that A) the HINT score post-rotation and minimization did decrease from -134 to -234, due to a new unfavorable interaction between ASP198's carboxylate and a nitrogen of the furazan ring and B) GOLD was unable to reproduce this pose, even with constraints. However, these HINT scores are still compared to a baseline of -385 of RocA, a known binder, in its co-crystallized pose, and, although we obtained many other high-scoring

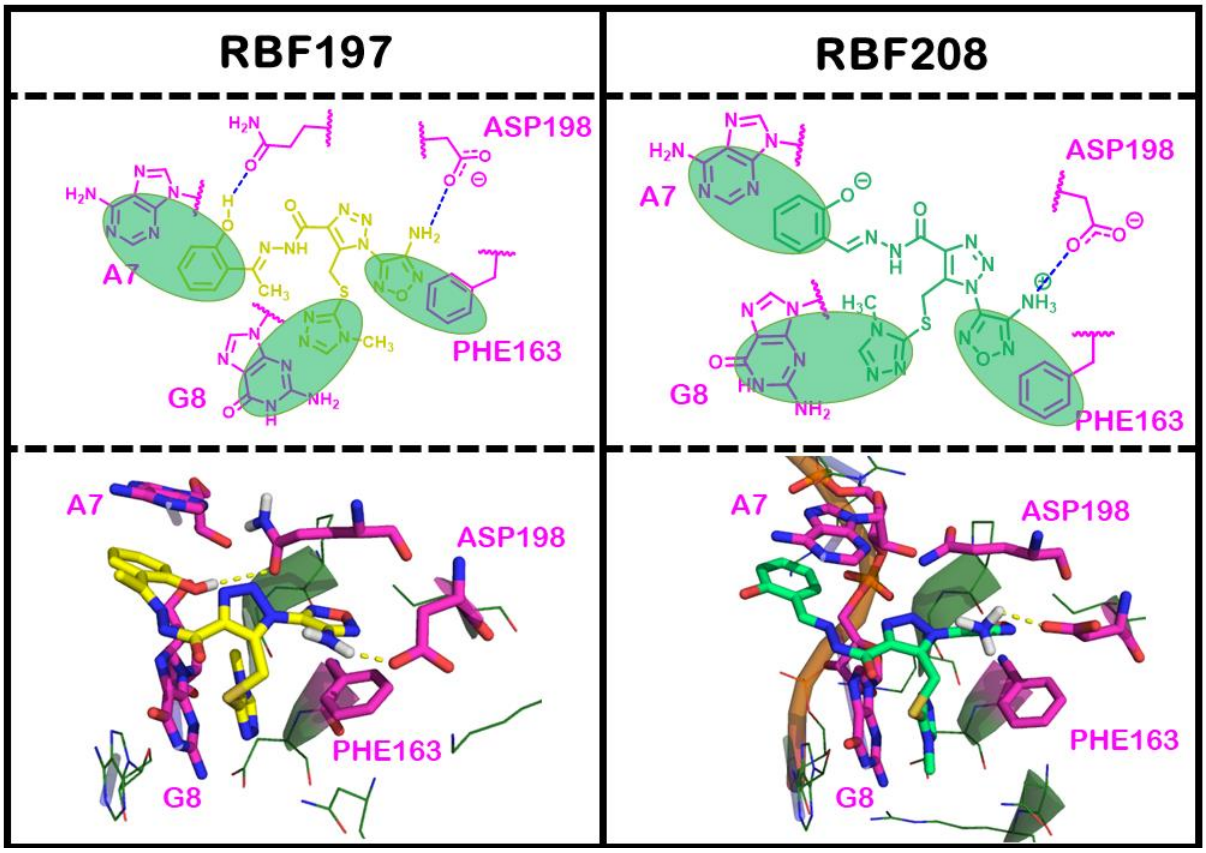


Figure 3.21. Docking poses of RBF197 and RBF208 in eIF4A:RNA groove. The top two panels show schematic representations of the interactions made between docked poses of RBF197 and RBF208 and their surrounding environments. Green, transparent ovals are used to two-dimensionally represent possible π-π stacking interactions between the ligands and surrounding residues. In these models, RBF197 and RBF208 both form π-π stacking interactions with the A7 and G8 bases and PHE163. Dashed lines between the ligands and surrounding residues are used to indicate hydrogen bonding, where the color indicates the donor/acceptor character of the ligand atom (blue = donor; red = acceptor). Both RBF197 and RBF208 form hydrogen bonding interactions with ASP198, but RBF197 donates an additional bond to GLN195. The lower two panels are high-scoring docked models of RBF197 and RBF208.

docked poses, we hypothesize that the poses of Figure 3.21 are highly probable, as their aromatic ring systems and interactions with GLN195 are consistent with our defined pharmacophore from the original virtual screen, which in turn was based on the RocA-bound crystal structure[43] of eIF4A1. Also, the pose shown for RBF197 was the most reasonable since this hit was notably more potent than the others, which we attribute to a crucial

hydrogen bond between the eIF4A1:RNA complex and the, in this case, protonated phenoxide moiety of our scaffold (Figures 3.9 and 3.21).

Our modeling studies suggested that RocA's and our compounds' mechanism of action involves trapping and distorting RNA's bound pose. Indeed, RocA's crystallized conformation positions itself such that it inserts between the A7 and G8 bases and binds on top of the bound RNA (Figure 3.4). From the results of our virtual screen, we believe our hit compounds bind similarly because they, too, have three aromatic rings capable of forming π-π stacking interactions (Figures 3.9 and 3.21). We speculated that such molecules trap the eIF4A:RNA complex. To experimentally confirm this, we employed a functional RNA unwinding assay to measure the activity of human eIF4A1 recombinant protein in the presence or absence of the inhibitors. Here, the RNA stable duplex was formed by annealing 32mer RNA modified with cyanine 5 (Cy5) at its 5′-end and a complementary 9mer modified with a cyanine 3 (Cy3) at the 3′-end (Supp Figure 9A).[34] A stable fluorescence was recorded. A 10-molar excess of unlabeled 9mer was added to the reaction to ensure a single turnover of the RNA unwinding. The reaction was started by adding excess ATP in the presence or absence of compounds. An increase in fluorescence readout was observed in the compound-treated samples compared to DMSO (this was considered basal) (Supp. Figure 9B).[34] To assess whether the compounds stably locked the eIF4A:RNA duplex, we added an additional fluorescent-labeled stable RNA complex and measured the kinetic values. The values were subsequently normalized and converted to percentage inhibition with respect to DMSO control groups. As proposed, the rate of RNA unwinding of the eIF4A:RNA duplex post-compound treatment was drastically reduced, with the IC_{50} value observed to be upper

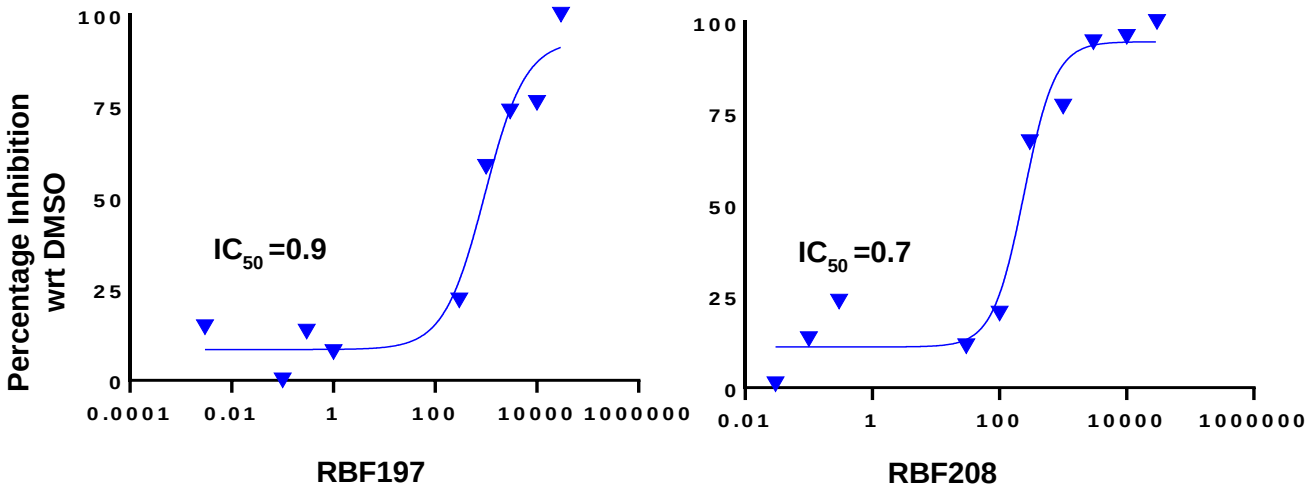


Figure 3.22. The difference in the increase in the fluorescence was calculated in the presence and absence of RBF197 and RBF208. Concentration-response curves were plotted using a graph pad prism and IC_{50} was observed at 0.7 and 0.9μM respectively.

nanomolar range (Figure 3.22). The assay results represent the potential RNA-eIF4A-inhibitor complex formation which diminishes eIF4A1's helicase activity.

Discussion

Translation initiation, particularly eIF4A RNA helicase, is emerging as a privileged chemotherapeutic target as numerous studies associate it with the rate of protein biosynthesis, tumor initiation, chemoresistance, cancer stem cell functions and metastasis[44-47]. While our current and previous studies[7,10,15,48] support the concept that eIF4A is critical in lymphomagenesis, the clinical utility of selective eIF4A inhibitors has been limited to date. Several potent small molecules have been identified, including natural molecules like rocaglates and elatol, demonstrating potent anticancer activity both *in vitro* and *in vivo*. However, none of these compounds have found success in the clinic.[9,18,49,50] An important exception is eFT226, a promising candidate undergoing Phase I clinical trials with the data still pending.[19] Resistance and relapse to frontline therapy in DLBCL still presents a major clinical issue. Therefore, the successful development of eIF4A-selective small molecules

inhibitors as a drug target, may open up new options for therapy of this most common adult lymphoma. Significantly, most potent eIF4A inhibitors, including eFT226, exhibit a common rocaglate backbone, raising the question of whether this chemical backbone is associated with the limiting toxicity. Furthermore, RocA is also reported to bind with prohibitin 1 and 2, thus impeding c-Raf induced MAPK/ERK pathways, raising the question about the specificity of this class of small molecules.[51]

We utilized the publicly available information about eIF4A1-RocA structure and designed a structure-guided approach to develop RocA-independent potent eIF4A small molecule inhibitors to address these shortcomings. We successfully identified three compounds, RBF197, RBF203, and RBF208, that hamper *in vitro* eIF4A helicase activity (IC_{50} ≤250pM). Furthermore, through selective replacement of a specific phenoxide moiety, we gained critical mechanistic insights into the mode of action for these small molecules to exert their inhibition. Lastly, we demonstrate that novel eIF4A inhibitors significantly hamper eIF4A-dependent target genes in biologically relevant DLBCL cells.

Our study began with a survey of the RocA binding site of a RocA:RNA:eIF4A1 co-crystallized complex. Using the HINT force field, we identified several key interactions that we attributed to RocA's tight binding to the RNA:eIF4A1 complex, including three π-π stacking and two different hydrogen-bonding interactions. These major interactions were utilized as features for a ligand pharmacophore-based virtual screen for novel eIF4A inhibitors. We obtained 1218 hits from our screen, which underwent high-throughput docking in GOLD using our original pharmacophore features as docking constraints. We purchased the top 29 best scoring compounds for primary screening. Based on the

previously established screening protocol targeting eIF4A1, we developed a luminometric method for screening eIF4A activity assay by measuring the eIF4A-sensitive 5'UTR driven luciferase readout.[12] This method is simple, sensitive, robust, and in-cell, providing quick and reliable outputs. Comparatively, all the other small molecules targeting eIF4A have been screened using *in vitro* assays.[9,52] Our preliminary screen noted RBF98 impeding 50% luciferase inhibition at 1 nM concentration in the luciferase-based assay while having minimal impact on control assays (Supp. Figure 3A).[34] Next, we ran it through an *in vitro* assay and observed that the compound inhibits ~50% of eIF4A helicase activity around 3 μM. We identified potential new interaction sites, like ASP198 and the known RocA binding sites of PHE163. This is important to note because our docking studies suggest that our most potent compounds from screen utilize similar features as rocaglates for binding and additional ones that may improve its selectivity for eIF4A1 and drug-like properties. The bountiful information we have obtained from our docking studies will further guide the design of new, more potent compounds.

Using these insights, we utilized a rational approach and searched for procurable chemical mimetics of RBF98 and identified three potent small molecule inhibitors: RBF197, RBF203, and RBF208. All three hits showed a remarkable selective decrease in luciferase assays. Notably, biochemical activity assays demonstrated compounds that are active at picomolar concentrations, which to our knowledge is the first report of eIF4A1 inhibitory activity at this potency. More importantly, all the three novel molecules displayed robust inhibition in cellular proliferation of DLBCLs with EC_{50} ranging in lower micromolar concentrations.

We next extended our study to delineate the mechanistic profiling of eIF4A-dependent transcripts in DLBCL using the MYC/BCL2 DLBCL cell line (RC[53]) and the ABC-DLBCL cell line (OCI-Ly3[54]). As anticipated, the compounds were effective in blocking translational output in DLBCL.[47] In fact, RBF197 and RBF208 showed a dose-dependent decrease in eIF4A-dependent oncogenes (cMYC,[55] MCL1,[54] and CARD11[56]). NRF2 is a redox master regulator induced by oncogenic KRas regulating the transcriptional program of specific translational factors for efficient protein synthesis.[57] Further, a recent report indicates that NRF2 activation, an emerging prognostic indicator in DLBCL,[58] confers resistance to silvestrol analog in cancer therapy.[57] In contrast, treatment with novel identified eIF4A inhibitors, we noted a dose-dependent decrease of NRF2. Similarly, CDK7, a critical cell cycle modulator deregulated in cMYC and BCL6 dependent DLBCL,[59] was also depleted upon the compound treatment. Likewise, PARP1, a DNA binding protein associated with DNA damage repair that confers resistance to genotoxic compounds routinely used as chemotherapeutic agents,[60] was also noted to decline in DLBCL cells. One of the major limitations of the previously reported eIF4A inhibitors like elatol[58] was unintended cytotoxicity under cellular studies. Thus, to define the therapeutic window, we performed the colony formation assays using malignant DLBCL cells and non-malignant transformed lymphoblastoid (GMO cell lines) cells. Surprisingly, RBF197 has a therapeutic edge over the counterpart RBF208 by showing the least toxic effect on the GMO cell colonies while inhibiting DLBCL colonies in a dose-dependent manner.

In silico analysis indicated the presence of a crucial hydrogen bond between the GLN195 of eIF4A1:RNA complex and the phenol moiety protonated, in this case, phenoxide moiety of our most active hit compound RBF197 (Figures 3.14B, 3.21), which potentially

explains its higher potency over our other hits. Notably, this docked pose of RBF197 forms the GLN195-phenol hydrogen bond with the acceptor end of GLN195's amide, which is different from RocA's hydrogen bond with the donor end. We also believe our docking suggests that our best hit compounds act by trapping eIF4A1 in an RNA-bound state. To address this, we performed an RNA trap assay to delineate the mechanism of action of the novel pharmacophores. This uncompetitive mechanism means that RBF197 or RBF208 prefers to bind eIF4A when the protein is in the RNA-bound state, which is advantageous, as binding RNA to eIF4A facilitates creating a clamp and leads to the unavailability of the enzyme for the next turnover cycle. This mechanism is more beneficial than the ATP inhibitors, resulting in unwanted toxicity in the cells.

Conclusion

Further work is required to develop a deeper understanding of this novel small molecule series and to describe in detail the mode of eIF4A inhibition. Additional molecular modeling studies coupled with medicinal chemistry design and synthesis of novel compounds will likely lead to even more efficacious lead compounds. As noted above, we have to date only explored analogues that vary the *m*-phenoxide moiety of RBF98, and significant scope for structure optimization is available in other regions of these molecules. Also, we are currently awaiting mutation study data related to PHE163, GLN195, ASP198 to determine the roles these residues play in binding our hit molecules, if any. Nevertheless, the compounds reported here are potent and unique structural eIF4A inhibitors forming different sets of interactions than the previously known inhibitors. These inhibitors have potential pharmacological relevance and present a valuable therapeutic opportunity.

Acknowledgments

This work was supported in part by a Merit Review Award from the Department of Veterans Affairs (RBG) and R01CA164311 (RBG) from the National Institutes of Health and shared resource and funding Massey NIH-NCI Cancer Center Support Grant P30 CA016059. The VCU School of Pharmacy and a Graduate School Dissertation Fellowship provided support for NBH.

References

1. Sanchez-Vega, F.; Mina, M.; Armenia, J.; Chatila, W. K.; Luna, A.; La, K. C.; Dimitriadoy, S.; Liu, D. L.; Kantheti, H. S.; Saghafinia, S.; Chakravarty, D.; Daian, F.; Gao, Q.; Bailey, M. H.; Liang, W.; Foltz, S. M.; Shmulevich, I.; Ding, L.; Heins, Z.; Ochoa, A.; Gross, B.; Gao, J.; Zhang, H.; Kundra, R.; Kandoth, C.; Bahceci, I.; Dervishi, L.; Dogrusoz, U.; Zhou, W.; Shen, H.; Laird, P. W.; Way, G. P.; Greene, C. S.; Liang, H.; Xiao, Y.; Wang, C.; Iavarone, A.; Berger, A. H.; Bivona, T. G.; Lazar, A. J.; Hammer, G. D.; Giordano, T.; Kwong, L. N.; McArthur, G.; Huang, C.; Tward, A. D.; Frederick, M. J.; McCormick, F.; Meyerson, M.; Cancer Genome Atlas Network; Van Allen, E. M.; Cherniack, A. D.; Ciriello, G.; Sander, C.; Schultz, N. Oncogenic Signaling Pathways in the Cancer Genome Atlas. *Cell* **2018**, *173*, 321-337.
2. Hagner, P. R.; Schneider, A; Gartenhaus, R. B. Targeting the translational machinery as a novel treatment strategy for hematologic malignancies. *Blood* **2010**, *115*, 2127-2135.
3. Truitt M. L.; Ruggero, D. New frontiers in translational control of the cancer genome. *Nat. Rev. Cancer* **2016**, *16*, 288-304.
4. Pelletier, J.; Graff, J.; Ruggero, D.; Sonenberg, N. Targeting the eIF4F translation initiation complex: a critical nexus for cancer development. *Cancer Res.* **2015**, *75*, 250-263.

5. Mitchell S.F.; Walker, S. E.; Algire, M. A.; Park, E.; Hinnebusch, A. G.; Lorsch, J. R. The 5'-7-methylguanosine cap on eukaryotic mRNAs serves both to stimulate canonical translation initiation and to block an alternative pathway. *Mol. Cell* **2010**, *39*, 950-962.
6. Lindqvist, L.; Imataka, H.; Pelletier, J. Cap-dependent eukaryotic initiation factor-mRNA interactions probed by cross-linking. *RNA* **2008**, *14*, 960-969.
7. Raza, F.; Waldron, J. A.; Quesne, J. L. Translational dysregulation in cancer: eIF4A isoforms and sequence determinants of eIF4A dependence. *Biochem. Soc. Trans.* **2015**, *43*, 1227-1233.
8. Xue, C.; Gu, X.; Li, G.; Bao, Z.; Li L. Expression and Functional Roles of Eukaryotic Initiation Factor 4A Family Proteins in Human Cancers. *Front. Cell. Dev. Biol.* **2021**, *9*, 711965.
9. Naineni S. K.; Maïga, R. I.; Cencic, R.; Putnam, A. A.; Amador, L. A.; Rodriguez, A. D.; Jankowsky, E.; Pelletier, J. A comparative study of small molecules targeting eIF4A. *RNA* **2020**, *26*, 541-549.
10. Andreou A.Z.; Klostermeier, D. (2013) The DEAD-box helicase eIF4A: paradigm or the odd one out? *RNA Biol.* **2013**, *10*, 19-32.
11. Rubio C. A.; Weisburd, B.; Holderfield, M.; Arias, C.; Fang, E.; DeRisi, J. L.; Fanidi, A. (2014) Transcriptome-wide characterization of the eIF4A signature highlights plasticity in translation regulation. *Genome Biol*. **2014**, *15*, 476.
12. Wolfe A. L.; Singh, K.; Zhong, Y.; Drewe, P.; Rajasekhar, V. K.; Sanghvi, V. R.; Mavrakis, K. J.; Jiang, M.; Roderick, J. E.; Van der Meulen, J.; Schatz, J. H.; Rodrigo, C. M.; Zhao, C.; Rondou, P.; de Stanchina, E.; Teruya-Feldstein, J.; Kelliher, M. A.; Speleman, F.; Porco Jr., J. A.; Pelletier, J.; Rätsch, G.; Wendel, H. RNA G-quadruplexes cause eIF4A-dependent oncogene translation in cancer. *Nature* **2014**, *513*, 65-70.

13. Modelska, A.; Turro, E.; Russell, R.; Beaton, J.; Sbarrato, T.; Spriggs, K.; Miller, J.; Gräf, S.; Provenzano, E.; Blows, F.; Pharoah, P.; Caldas, C.; Quesne, J. L. The malignant phenotype in breast cancer is driven by eIF4A1-mediated changes in the translational landscape. Cell Death Dis. **2017**, *6*, e1603.
14. Li, W.; Chen, A.; Xiong, L.; Chen, T.; Tao, F.; Lu, Y.; He, Q.; Zhao, L.; Ou, R.; Xu, Y. (2017) miR-133a acts as a tumor suppressor in colorectal cancer by targeting eIF4A1. *Tumour Biol*. **2017**, *39*, 1010428317698389.
15. Liang, S.; Zhou, Y.; Chen, Y.; Ke, G.; Wen, H.; Wu, X. Decreased expression of EIF4A1 after preoperative brachytherapy predicts better tumor-specific survival in cervical cancer. *Int. J. Gynecol. Cancer* **2014**, *24*, 908-915.
16. Gao, C.; Guo, X.; Xue, A.; Ruan, Y.; Wang, H.; Gao, X. High intratumoral expression of eIF4A1 promotes epithelial-to-mesenchymal transition and predicts unfavorable prognosis in gastric cancer. *Acta. Biochim. Biophys. Sin. (Shanghai)*, **2020**, *52*, 310-319.
17. Cencic, R.; Pelletier, J. Hippuristanol - A potent steroid inhibitor of eukaryotic initiation factor 4A. *Translation (Austin)* **2016**, *4*, e1137381.
18. Chu, J.; Zhang, W.; Cencic, R.; Devine, W. G.; Beglov, D.; Henkel, T.; Brown, L. E.; Vajda, S.; Porco, Jr., J. A.; Pelletier, J. Amidino-Rocaglates: A Potent Class of eIF4A Inhibitors. *Cell Chem. Biol.* **2019**, *26*, 1586-1593.
19. Ernst, J. T.; Thompson, P. A.; Nilewski, C.; Sprengeler, P. A.; Sperry, S.; Packard, G.; Michels, T.; Xiang, A.; Tran, C.; Wegerski, C. J.; Eam, B.; Young, N. P.; Fish, S.; Chen, J.; Howard, H.; Staunton, J.; Molter, J.; Clarine, J.; Nevarez, A.; Chiang, G. G.; Appleman, J. R.; Webster, K. R.; Reich, S. H. Design of Development Candidate eFT226, a First in Class Inhibitor of Eukaryotic Initiation Factor 4A RNA Helicase. *J. Med. Chem.* **2020**, *63*, 5879-5955.

20. Berman H. M.; Westbroook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.;Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res*. **2000**, *28*, 235-242.
21. Jones, G.; Willett, P.; Glen, R. C.; Leach, A. R.; Taylor, R. Development and Validation of a Genetic Algorithm for Flexible Docking. *J. Mol. Biol.* **1997**, *267*, 727-748.
22. Kellogg, G. E.; Semus, S. F.; Abraham, D. J. HINT: A New Method of Empirical Hydrophobic Field Calculation for CoMFA. *J. Comput. Aided Mol. Des.* **1991**, *5*, 545-552.
23. Kellogg G. E.; Abraham, D. J. Hydrophobicity: is LogP(o/w) More Than the Sum of its Parts? *Eur. J. Med. Chem.* **2000**, *35*, 651-661.
24. Chandrashekar D.S.; Bashel, B.; Balasubramanya, S. A. H.; Creightoon, C. J. Ponce-Rodriguez, I.; Chakravarthi, B. V. S. K.; Varambally, S. UALCAN: A Portal for Facilitating Tumor Subgroup Gene Expression and Survival Analyses. *Neoplasia* **2017**, *19*, 649-658.
25. Chandrashekar, D. S.; Karthikeyan, S. K.; Korla, P. K.; Patel, H.; Shovon, A. R.; Athar, M.; Netto, G. J.; Qin, Z. S.; Kumar, S.; Manne, U.; Creighton, C. J.; Varambally, S. UALCAN: An update to the integrated cancer data analysis platform. *Neoplasia* **2022**, *25*, 18-27.
26. Schmiedel, B. J.; Singh, D.; Madrigal, A.; Valdovino-Gonzalez, A. G.; White, B. M.; Zapardiel-Gonzalo, J.; Ha, B.; Altay, B.; Greenbaum, J. A.; McVicker, G.; Seumois, G.; Rao, A.; Kronenberg, M.; Peters, B.; Vijayanand, P. Impact of Genetic Polymorphisms on Human Immune Cell Gene Expression. *Cell* **2018**, *175*, 1701-1715.
27. Schmitz, R.; Wright, G. W.; Huang, D. W.; Johnson, C. A.; Phelan, J. D.; Wang, J. Q.; Roulland, S.; Kasbekar, M.; Young, R. M.; Shaffer, A. L.; Hodson, D. J.; Xiao, W.; Yu, X.; Yang, Y.; Zhao, H.; Xu, W.; Liu, X.; Zhou, B.; Du, W.; Chan, W. C.; Jaffe, E. S.; Gascoyne, R. D.; Connors, J. M.; Campo, E.; Lopez-Guillermo, A.; Rosenwald, A.; Ott, G.; Delabie, J.; Rimsza, L. M.; Wei, K. T. K.; Zelenetz, A. D.; Leonard, J. P.; Bartlett, N. L.; Tran, B.; Shetty, J.; Zhao, Y.; Soppet, D. R.;

Pittaluga, S.; Wilson, W. H.; Staudt, L. M. Genetics and Pathogenesis of Diffuse Large B-Cell Lymphoma. *N. Engl. J. Med*. **2018**, *378*, 1396-1407.

28. Lenz, G.; Wright, G.; Dave, S. S.; Xiao, W.; Powell, J.; Zhao, H.; Xu, W.; Tan, B.; Goldschmidt, N.; Iqbal, J.; Vose, J.; Bast, M.; Fu, K.; Weisenberger, D. D.; Greiner, T. C.; Armitage, J. O.; Kyle, A.; May, L.; Gascoyne, R. D.; Connors, J. M.; Troen, G.; Holte, H.; Kvaloy, S.; Dierickx, D.; Verhoef, G.; Delabie, J.; Smeland, E. B.; Jares, P.; Martinez, A.; Lopez-Guillermo, A.; Montserrat, E.; Campo, E.; Braziel, R. M.; Miller, T. P.; Rimsza, L. M.; Cook, J. R.; Pohlman, B.; Sweetenham, J.; Tubbs, R. R.; Fisher, R. I.; Hartmann, E.; Rosenwald, A.; Ott, G.; Muller-Hermelink, H.; Wrench, D.; Lister, T. A.; Jaffe, E. S.; Wilson, W. H.; Chan, W. C.; Staudt, L. M.; Lymphoma/Leukemia Molecular Profiling Project. Stromal Gene Signatures in Large-B-Cell Lymphomas. *N. Engl. J. Med.* **2008**, *359*, 2313-2323.
29. Cardesa-Salzmann, T. M.; Colomo, L.; Gutierrez, G.; Chan, W. C.; Weisenberger, D.; Climent, F.; González-Barca, E.; Mercadal, S.; Arenillas, L.; Serrano, S.; Tubbs, R.; Delabie, J.; Gascoyne, R. D.; Connors, J. M.; Mate, J. L.; Rimsza, L.; Braziel, R.; Rosenwald, A.; Lenz, G.; Wright, G.; Jaffe, E. S.; Staudt, L.; Jares, P.; López-Guillermo, A.; Campo, E. High Microvessel Density Determines a Poor Outcome in Patients With Diffuse Large B-Cell Lymphoma Treated With Rituximab Plus Chemotherapy. *Haematologica* **2011**, *96*, 996-1001.
30. Dubois, S.; Viailly, P.; Bohers, E.; Betrand, P.; Ruminy, P.; Marchand, V.; Maingonnat, C.; Mareschal, S; Picquenot, J.; Penther, D.; Jais, J.; Tesson, B.; Peyrouze, P.; Figeac, M.; Desmots, F.; Fest, T.; Haioun, C.; Lamy, T.; Copie-Bergman, C.; Fabiani, B.; Delarue, R.; Peyrade, F.; André, M.; Ketterer, N.; Leroy, K.; Salles, G.; Molina, T. J.; Tilly, H.; Jardin, F. Biological and Clinical Relevance of Associated Genomic Alterations in MYD88 L265P and

non-L265P-Mutated Diffuse Large B-Cell Lymphoma: Analysis of 361 Cases. *Clin. Cancer Res.* **2017**, *23*, 2232-2244.

31. Dubois, S.; Tesson, B.; Mareschal, S.; Viailly, P.; Bohers, E.; Ruminy, P.; Etancelin, P.; Peyrouze, P.; Copie-Bergman, C.; Fabiani, B.; Petrella, T.; Jais, J.; Haioun, C.; Salles, G.; Molina, T. J.; Leroy, K.; Tilly, H.; Jardin, F.; Lymphoma Study Association (LYSA) Investigators. Refining diffuse large B-cell lymphoma subgroups using integrated analysis of molecular profiles. *EBioMedicine* **2019**, *48*, 58-69.
32. Menon, M. P.; Pittaluga, S.; Jaffe, E. S. The Histological and Biological Spectrum of Diffuse Large B-Cell Lymphoma in the World Health Organization Classification. *Cancer J*. **2012**, *18*, 411-420.
33. Nowakowski, G. S.; Czuczman, M. S. ABC, GCB, and Double-Hit Diffuse Large B-Cell Lymphoma: Does Subtype Make a Difference in Therapy Selection? *Am. Soc. Clin. Oncol. Educ. Book* **2015**, e449-457.
34. Kayastha, F.; Herrington, N. B.; Kapadia, B. Roychowdhury, A.; Nanaji, N.; Kellogg, G. E.; Gartenhaus, R. B. Novel eIF4A1 Inhibitors with Anti-Tumor Activity in Lymphoma. *Molecular Medicine*, submitted.
35. Iwasaki, S.; Iwasaki, W.; Takahashi, M.; Sakamoto, A.; Watanabe, C.; Shichino, Y.; Floor, S. N.; Fujiwara, K.; Mito, M.; Dodo, K.; Sodeoka, M.; Imataka, H.; Honma, T.; Fukuzama, K.; Ito, T.; Ingolia, N. T. The Translation Inhibitor Rocaglamide Targets a Bimolecular Cavity between eIF4A and Polypurine RNA. *Mol. Cell* **2019**, *73*, 738-748.
36. Obaidullah, A. J.; Ahmed, M. H.; Kitten, T.; Kellogg, G. E. Inhibiting Pneumococcal Surface Antigen A (PsaA) With Small Molecules Discovered Through Virtual Screening: Steps

Toward Validating a Potential Target for Streptococcus Pneumoniae. *Chem. Biodivers*. **2018**, *15*, e1800234.

37. Spyrakis, F.; Singh, R.; Cozzini, P.; Campanini, B.; Salsi, E.; Felici, P.; Raboni, S.; Benedetti, P.; Cruciani, G.; Kellogg, G. E.; Cook, P. F.; Mozzarelli, A. Isozyme-Specific Ligands for O-Acetylserine Sulfhydrylase, a Novel Antibiotic Target. *PLoS One* **2013**, *8*, e77558.
38. Chen, D.; Misra, M.; Sower, L.; Peterson, J. W.; Kellogg, G. E.; Schein, C. H. Novel Inhibitors of Anthrax Edema Factor. *Bioorg. Med. Chem.* **2008**, *16*, 7225-7233.
39. Hurst, T. Flexible 3D Searching: The Directed Tweak Technique. J. Chem. Inf. Comp Sci. **1994**, *34*, 190-196.
40. Zhang, X.; Bi, C.; Lu, T.; Zhang, W.; Yue, T.; Wang, C.; Tian, T.; Zhang, X.; Huang, Y.; Lunning, M.; Hao, X.; Brown, L. E.; Devine, W. G.; Vose, J.; Porco, Jr., J. A.; Fu, K. Targeting Translation Initiation by Synthetic Rocaglates for Treating MYC-Driven Lymphomas. *Leukemia* **2020**, *34*, 138-150.
41. Stoneley, M.; Willis, A. E. eIF4A1 is a Promising New Therapeutic Target in ER-Negative Breast Cancer. *Cell Death Differ.* **2015**, *22*, 524-525.
42. Rogers, D. J.; Tanimoto, T. T. A Computer Program for Classifying Plants. *Science* **1960**, *132*, 1115-1118.
43. Juskevicius, D.; Müller, A.; Hashwah, H.; Lundberg, P.; Tzankov, A.; Menter, T. Characterization of the Mutational Profile of 11 Diffuse Large B-Cell Lymphoma Cell Lines. *Leuk. Lymphoma* **2018**, *59*, 1710-1716.
44. Fabbri, L.; Chakraborty, A.; Robert, C.; Vagner, S. The Plasticity of mRNA Translation During Cancer Progression and Therapy Resistance. *Nat. Rev. Cancer* **2021**, *21*, 558-577.

45. Park, E.; Wang, Q.; Thakar, M.; Ren, Z.; Soori, M.; Gondek, L.; Matsui, W.; Gocke, C. Protein Synthesis Rates Regulate Tumor-initiating Potential and Chemoresistance in Multiple Myeloma. Clinical Lymphoma, Myeloma & Leukemia **2019**, *19*, e129.
46. Lee L. J.; Papadopoli, D.; Jewer, M.; Del Rincon, S.; Topisirovic, I.; Lawrence, M. G.; Postovit, L. Cancer Plasticity: The Role of mRNA Translation. *Trends Cancer* **2021**, *7*, 134-145.
47. Chan, K.; Robert, F.; Oertlin, C.; Kapeller-Libermann, D.; Avizonis, D.; Gutierrez, J.; Handly-Santana, A.; Doubrovin, M.; Park, J.; Schoepfer, C.; Da Silva, B.; Yao, M.; Gorton, F.; Shi, J.; Thomas, C. J.; Brown, L. E.; Porco, Jr., J. A.; Pollak, M.; Larsson, O.; Pelletier, J.; Chio, I. I. C. eIF4A Supports an Oncogenic Translation Program in Pancreatic Ductal Adenocarcinoma. *Nat. Commun.* **2019**, *10*, 5151.
48. Kapadia, B.; Nanaji, N. M.; Bhalla, K.; Bhandary, B.; Lapidus, R.; Beheshti, A.; Evens, A. M.; Gartenhaus, R. B. Fatty Acid Synthase Induced S6Kinase Facilitates USP11-eIF4B Complex Formation for Sustained Oncogenic Translation in DLBCL. *Nat. Commun.* **2018**, *9*, 829.
49. Peters, T. L.; Tillotson, J.; Yeomans, A. M.; Wilmore, S.; Lemm, E.; Jiménez-Romero, C.; Amador, L. A.; Li, L.; Amin, A. D.; Pongtornpipat, P.; Zerio, C. J.; Ambrose, A. J.; Paine-Murrieta, G.; Greninger, P.; Vega, F.; Benes, C. H.; Packham, G.; Rodríguez, A. D.; Chapman, E.; Schatz, J. H. Target-Based Screening against eIF4A1 Reveals the Marine Natural Product Elatol as a Novel Inhibitor of Translation Initiation with In Vivo Antitumor Activity. *Clin. Cancer Res.* **2018**, *24*, 4256-4270.
50. Chu, J.; Pelletier, J. Therapeutic Opportunities in Eukaryotic Translation. *Cold Spring Harb. Perspect. Biol*. **2018**, *10*, a032995.

51. Chu, J.; Galicia-Vázquez, G.; Cencic, R.; Mills, J. R.; Katigbak, A.; Porco, Jr., J. A.; Pelletier, J. CRISPR-Mediated Drug-Target Validation Reveals Selective Pharmacological Inhibition of the RNA Helicase, eIF4A. *Cell Rep*. **2016**, *15*, 2340-2347.
52. Abdelkrim, Y. Z.; Harigua-Souiai, E.; Barhoumi, M.; Banroques, J.; Blondel, A.; Guizani, I.; Tanner, N. K. The Steroid Derivative 6-Aminocholestanol Inhibits the DEAD-Box Helicase eIF4A (LieIF4A) from the Trypanosomatid Parasite Leishmania by Perturbing the RNA and ATP Binding Sites. *Mol. Biochem. Parasitol.* **2018**, *226*, 9-19.
53. Pham, L. V.; Lu, G.; Tamayo, A. T.; Chen, J.; Challagundla, P.; Jorgensen, J. L.; Medeiros, L. J.; Ford, R. J. Establishment and Characterization of a Novel MYC/BCL2 "Double-Hit" Diffuse Large B Cell Lymphoma Cell Line, RC. *J. Hematol. Oncol*. **2015**, *8*, 121.
54. Wenzel, S.; Grau, M.; Mavis, C.; Hailfinger, S.; Wolf, A.; Madle, H.; Deeb, G.; Dörken, B.; Thome, M.; Lenz, P.; Dirnhofer, S.; Hernandez-Ilizaliturri, F. J.; Tzankov, A.; Lenz, G. MCL1 is Deregulated in Subgroups of Diffuse Large B-Cell Lymphoma. *Leukemia* **2013**, *27*: 1381-1390.
55. Wilmore, S.; Rogers-Broadway, K.; Taylor, J.; Lemm, E.; Fell, R.; Stevenson, F. K.; Forconi, F.; Steele, A. J.; Coldwell, M.; Packham, G.; Yeomans, A. Targeted Inhibition of eIF4A Suppresses B-cell Receptor-Induced Translation and Expression of MYC and MCL1 in Chronic Lymphocytic Leukemia Cells. *Cell Mol. Life Sci.* **2021**, *78*, 6337-6349.
56. Steinhardt JJ*, et al.* (2014) Inhibiting CARD11 translation during BCR activation by targeting the eIF4A RNA helicase. *Blood* **124:** 3758-3767.
57. Chio, I. I. C.; Jafarnejad, S. M.; Ponz-Sarvise, M.; Park, Y.; Rivera, K.; Palm, W.; Wilson, J.; Sangar, V.; Hao, Y.; Öhlund, D.; Wright, K.; Filippini, D.; Lee, E. J.; Da Silva, B.; Schoepfer, C.; Wilkinson, J. E.; Buscaglia, J. M.; DeNicola, G. M.; Tiriac, H.; Hammell, M.; Crawford, H. C.;

Schmidt, E. E.; Thompson, C. B.; Pappin, D. J.; Sonenberg, N.;Tuveson, D. A. NRF2 Promotes Tumor Maintenance by Modulating mRNA Translation in Pancreatic Cancer. Cell **2016**, *166*, 963-976.

58. Yi, X.; Zhao, Y.; Xue, L.; Zhang, J.; Qiao, Y.; Jin, Q.; Li, H. Expression of Keap1 and Nrf2 in Diffuse Large B-Cell Lymphoma and its Clinical Significance. *Exp. Ther. Med*. **2018**, *16*, 573-578.
59. Lacrima, K.; Rinaldi, A.; Vignati, S.; Martin, V.; Tibiletti, M. G.; Gaidano, G.; Catapano, C. V.; Bertoni, F. Cyclin-Dependent Kinase Inhibitor Seliciclib Shows In Vitro Activity in Diffuse Large B-Cell Lymphomas. *Leuk. Lymphoma* **2007**, *48*, 158-167.
60. Hu, Y.; Lin, J.; Fang, H.; Fang, J.; Li, C.; Chen, W.; Liu, S.; Ondrejka, S.; Gong, Z.; Reu, F.; Maciejewski, J.; Yi, Q.; Zhao, J. Targeting the MALAT1/PARP1/LIG3 Complex Induces DNA Damage and Apoptosis in Multiple Myeloma. *Leukemia* **2018**, *32*, 2250-2262.

Chapter 4: Development of a Protein-Protein Interface Optimization Tool Using Hydropathic Environment Maps

Introduction

Harkening back to the subject of protein structure prediction tools, one area of interest in protein structure prediction is simulating the union of two proteins in so-called protein-protein docking. Surface residues on proteins, by nature, are highly flexible while they are immersed in solvent, but as they are brought in close proximity to the surface of another interacting protein, they will adopt more static conformations that are favorable to the formation of a protein-protein complex. The greatest reason for developing computational tools for predicting protein-protein complexes is that crystallization of multimeric complexes is significantly more expensive and time-consuming than for monomers.[1] Consequently, mutation studies, application of structure-based drug discovery approaches, etc. are similarly challenging and often not accessible. Thus, there is an unmet need for studying protein-protein interactions (PPI) and identifying new, potential druggable sites for new therapies. Designing inhibitors of PPIs presents potential new advantages (and challenges) compared to traditional *in-silico* orthosteric or even allosteric binding site drug discovery.[2] For one, PPI inhibitors can be designed to have greater selectivity for a target based on structural features of an interacting pair to overcome promiscuity associated with many other drug molecules. With this increased selectivity comes the beneficial secondary effect of not inhibiting downstream processes of multiple protein complexes. Challenges, however, include the fact that potential PPI sites may be too shallow to effectively conduct *in-silico* drug discovery efforts, as the formation of protein-

protein interfaces often form two largely planar surfaces between both structures. Additionally, it can prove difficult to construct effective functional assays to probe interactions at a buried PPI site. Nevertheless, design of robust PPI simulation tools presents a unique challenge with many rewards to reap that may be more than adequately met by computational techniques.

Computational studies of PPIs face significant challenges, not the least of which is first developing or identifying a proper energy scoring function that accurately assess the favorability and stability of a protein-protein complex and factors in considerations of newly formed hydrogen bonds and the hydrophobic effect (See Chapter 1).[3,4] Of course, this is a challenge that computational chemists and structural bioinformaticists have been tackling since the dawn of this field, but it is no less important to ensure that the force field implemented for this problem adequately simulates the free energy of the complex formation. For this reason, it is also important that the force field accurately simulates the free energies of solvation and desolvation, as water has been shown to play a pivotal role in the formation of a protein interface,[5-7] which our lab has also studied.[8,9] This role of solvation also raises a question of how to deal with protonation states of ionizable residues, since protonation can drastically change unfavorable interactions into favorable ones and vice versa.

Many labs have already undertaken this task and produced variations of protein-protein docking programs. An early protein-protein docking program was developed as an implementation of Rosetta, known as RosettaDock, by Gray et al.,[10] which uses a Monte Carlo-based method to generate solutions from a low-resolution docking, refine those solutions

using an interface optimization protocol, and then make predictions based on clustered solutions. HDOCK, developed by Yan et al.[11] uses a Fast-Fourier Transform (FFT)-based search method to compute putative solutions to docking two proteins and then evaluates solutions using an embedded scoring function. ClusPro, developed by Kozakov et al.,[12] uses a similar FFT-based method to perform a rigid body docking, sample billions of PPI conformations, cluster the 1000 lowest energy solutions by RMSD, and energy minimize the top solutions to remove steric clashes. Our lab has particularly dealt with ZDOCK and shown that explicit inclusion of water molecules in input files significantly improves docking accuracy of two proteins with this program.[13] Numerous others have been or are currently under development to address this problem from different perspectives.

Our ongoing 3D Interaction Homology (3DIH) project has many applications, among them being the development of protein structure prediction. We believe that a tool built to exploit these unique data and algorithms can predict the structures of two interacting proteins via a protein docking program, which we have made progress in developing. To illustrate how this may be done, it is important to recall that the objective of our 3DIH project is to exhaustively determine all of the hydropathic environments that can surround any given residue. With a number of reports that have studied some of these residues in-depth,[14-18] we have completed our mission of extracting this data from our data set for each residue and have constructed a library of hydropathic interaction maps detailing the hydropathic valences needed to "satiate" them. In theory, these maps represent the interactions each residue type "wants" to make with its environment, including the residues of incoming protein surfaces. A tool of our design should, therefore, predict how residues at a protein-

protein interface will mesh and interact together based on the most favorable combination of overlapping maps.

Methods and Results

Development of a Hydropathic Map Library

The most necessary precursor to constructing our protein-protein docking tool is to compile a library of hydropathic maps. These maps are intended to capture information about the kinds of interactions each residue type "wants" to make with its environment, what we term its "hydropathic valence." Our maps are constructed in the form of three-dimensional constellations of 4 separate map types (positive/negative hydrophobic and positive/negative polar) that indicate the position and magnitude of interactions each residue type forms with its environment. For more information on the data collection for these maps, see Chapter 2 of this thesis or any of our publications that have focused on particular residues.[14-18] To briefly reiterate, interaction data is collected by scoring interactions between a residue and its environment using the HINT force field,[19,20] which uses a dictionary of atomistic and residue-specific partial $\log P_{o/w}$ values encoding crucial free energy terms to score interactions between hydropathic species. This map data collection is applied to every residue of every type from a data set of 2703 protein structures selected from the Protein Data Bank.[21] The maps we collect are organized first by chess square, indicative of the residue's secondary structure, followed by the parse of its χ_1 dihedral angle into one of three groups (60°, 180°, and 300°) as in those of typical rotamer libraries,[22] and finally its cluster ID. Residues in the same cluster are represented by the same structure and map generated as averages from those residues composing the cluster. After

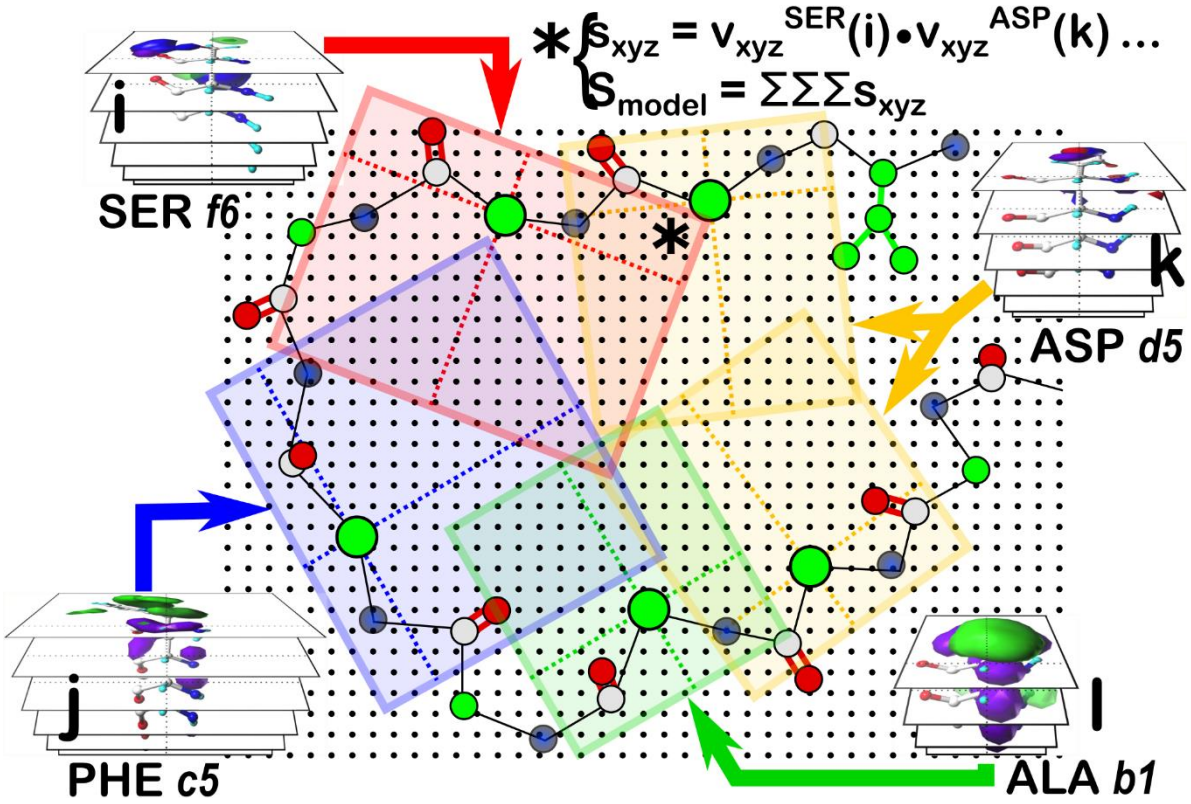


Figure 4.1. An illustration of a potential protein-protein interface and our designed scoring method. In this two-dimensional representation of a protein-protein docking, a sample of maps from our hydropathic map library are interpolated onto a master two-dimensional grid of points, spaced 0.5 Å apart, where each map is interpolated onto each interface residue at its C-alpha carbon. The value of each map at each master grid point is calculated through a series of geometric relations. The overall score of the protein-protein docking solution model in the pictured situation is calculated as a pairwise sum of the products of overlapping map values at each master grid point. For our purposes, this scoring system is translated into three dimensions.

compiling this data, we can more effectively begin constructing a variety of protein structure prediction tools, including one for protein-protein docking.

Model Construction and Scoring Philosophy

As a secondary step toward developing a protein-protein docking program, we first wish to complete a protein-protein interface residue optimization program, adopting the same philosophy described above. Our goal with this project is to design a program that will consider all residues at a protein-protein interface, perform map selection for residues at the interface, and optimize the positions of residue sidechains in such a way that they form the

most favorable interactions and reproduce the original crystal structure. We have developed our scoring system based on the favorable overlap of contours encoded into each of our maps, according to Figure 4.1. Our scoring system depends on constructing of a three-dimensional grid of points, spaced 0.5 Å apart and encompassing all residues at a protein-protein interface, where scoring takes place on a point-by-point basis. The overall model score, S_{model}, is calculated as a pairwise sum of multiplied overlapping map values from two interacting residues. Our model also leaves room for applying weighted consideration to each combination of map types. This strength of our approach is designed such that we have the potential to explore the impact of various combinations of map types (i.e., positive-hydrophobic-positive hydrophobic, positive-hydrophobic-negative hydrophobic, positive hydrophobic-positive polar, etc.) and determine the extent to which the interplay between these interaction types yields a better overall docked model.

Reframing Maps onto the Same Coordinate System

An important part of our methodology involves translating our map data into the master three-dimensional space, described above, where individual maps can interact with each other. For details on our map construction and format, see any of our communications regarding particular residues[14-18] or Chapter 2 of this thesis. Briefly, our maps individually occupy isolated space in Cartesian space, where every C-alpha carbon sits at the origin, the CA-CB bond is aligned to the Z-axis, and the CA-CH bond is oriented into the -y, -z quadrant of yz-space. In order to score map combinations together, each map grid point must be interpolated onto the same three-dimensional master grid system with all other scored maps. Also, since each individual map's grid points are unlikely to overlap with the master

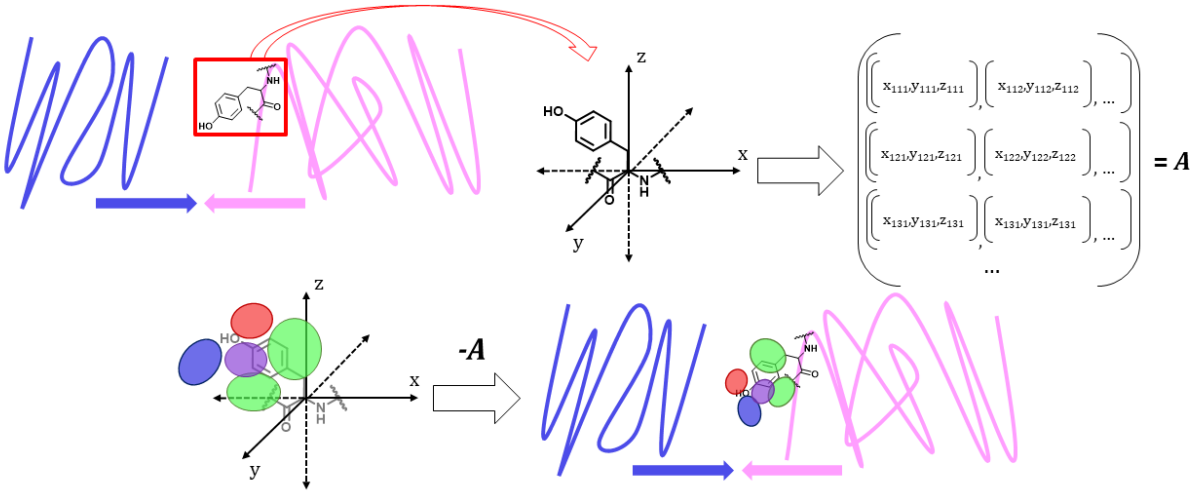


Figure 4.2. Diagram describing our map interpolation process. Beginning with a residue at a protein-protein interface, an orientation matrix is calculated based on three-dimensional movements required to orient the residue at the origin of Cartesian space. The negative form of this matrix is applied to a selected residue map, which already is oriented about the origin, in order to re-orient it toward the residue at the interface.

grid system's points, the specific values of each map grid point must be scaled, based on their proximity to known values at known points in space following map-to-master grid translation. To accomplish this, our program first constructs the master grid system. Then, residue-by-residue, we calculate an orientation matrix that moves and orients the residue to the origin, according to our standard map construction. The negative form of this calculated orientation matrix is then applied to re-orient each map to its residue on the master grid system, where all individual map grid points now are assigned known coordinates. Geometric calculations are performed to determine the hydropathic map values of in-between points of the master grid system, which can then participate in interaction scoring with other residue maps. A diagram of this operation can be seen in Figure 4.2. This approach enables us to score interacting maps in the same space, which is required for scoring any generated docking solution.

Scoring Function

In order to create a properly working protein-protein docking program, we must incorporate a properly designed fitness or scoring function that accurately assesses the success of a docking solution. As described (see Figure 4.1), we have designed our scoring function to be a pairwise sum of map values at every point on our master grid system. Since our maps are divided into a quartet of interaction types, all four map types must be considered at each grid point. There are a number of different features of our method undergoing optimization. For example, we have yet to determine the relative importance of each interaction type. Likewise, different combinations of interaction types may ultimately contribute more to scoring than others. For this reason, we intend to consider weighting certain terms from our scoring as we discover them to have more or less bearing on score output. It will be interesting to see the degree to which pure terms (i.e., positive polar-positive polar, positive hydrophobic-positive hydrophobic, etc.) and cross terms (i.e., positive polar-negative polar, positive polar-positive hydrophobic, etc.) will matter, as we expect some terms to have higher contribution to scoring and others to possibly be negligible. Certainly, trial-and-error is an option for determining relative weighting of each term, but a likely more effective solution would be multi-linear regression analysis to determine an overall equation of best fit for our function. The basic model of this equation (similar to S_{xyz} and S_{Model} in Figure 4.1) to calculate a score between two maps at each point, summed together, will look like:

$$S_{xyz} = AV(PP \cdot PP)_{xyz} + BV(NP \cdot NP)_{xyz} + CV(PH \cdot PH)_{xyz} + DV(NH \cdot NH)_{xyz} + EV(PP \cdot NP)_{xyz} + FV(PP \cdot PH)_{xyz} + \ldots$$

$$S_{Model} = \Sigma S_{xyz} ,$$

where S_{xyz} is the score at any point on the master grid system, V is the value of a particular scoring term, PP, NP, PH [and NH] are positive-polar, negative polar, positive hydrophobic [and negative hydrophobic] terms, respectively, A – E are relative weights of each pure/cross term, and S_{Model} is the total score of the model. This, like many aspects of our method to be described, will require significant experimentation to refine and implement.

Implementation of a Genetic Algorithm

The most challenging part of our approach is the vast quantity of map data we have generated for our library. For a typical residue, such as aspartic acid, its hydropathic map data is first divided by chess square, according to our chessboard schema, and then parsed by its χ_1 angle into three parses. For our purposes, we wish to construct a flexible docking program that varies the conformation of interface residue side chains but keeps its backbone dihedral angles constant. With this being the case, the chess squares for all residues of any input structure would be easily determinable and never change in our optimization algorithm. Where every chess square may contain up to three parses with up to 12 clusters per parse, any protein-protein docking may roughly be reduced to 36^n solutions, where *n* equals the number of residues at an interface. It should be noted that not every chess square contains residues in every parse or residues at all; however, this is counter-balanced by some residues having three additional subparses about their χ_2 angles (such as glutamic acid) and/or more than 12 clusters in each chess square (such as arginine). This results in an immense combinatorial problem that cannot realistically be solved by exhaustively evaluating all possible docking solutions. Rather, it should have some method for more

effectively sampling solutions more likely to be more energetically favorable. A number of methods exist for enhanced sampling to accomplish this. We have elected to implement a genetic algorithm for our purposes as our first approach to this problem.

A genetic algorithm (GA) is essentially designed to simulate the principles of natural selection to find the solution to a problem. The GA first performs a random selection of potential solutions to the problem, known as a population, evaluates the fitness of each solution, known as a chromosome/genome, and ranks them. The chromosomes with the highest fitness influence the production of a new generation of chromosomes, where, in theory, small variations of the best-performing chromosomes will create even higher performing ones. Certain chromosomes will undergo "crossover" together and produce offspring via exchange of their features. Other chromosomes will undergo "mutations" where their individual features can be randomly swapped out for other possible features to produce new "children." Some GAs also incorporate a practice of "elitism," which is designed to ensure that a certain number of the absolute best scoring chromosomes from every generation progresses to the following generation. These operational components of GAs are visually summarized in Figure 4.3. Many of these crossover and mutation events happen after evaluating a single population of possible chromosomes, producing offspring that become the start of a new generation of chromosomes. GAs run over a multitude of generations, sometimes several thousand or tens of thousands, in an attempt to reach the highest possible performing solution. The alternative approach would be to begin the GA with many thousands of trial parents and fewer generations, but the disadvantage of this would be too widespread sampling of potentially low-quality solutions with not enough generations of natural selection for refinement.

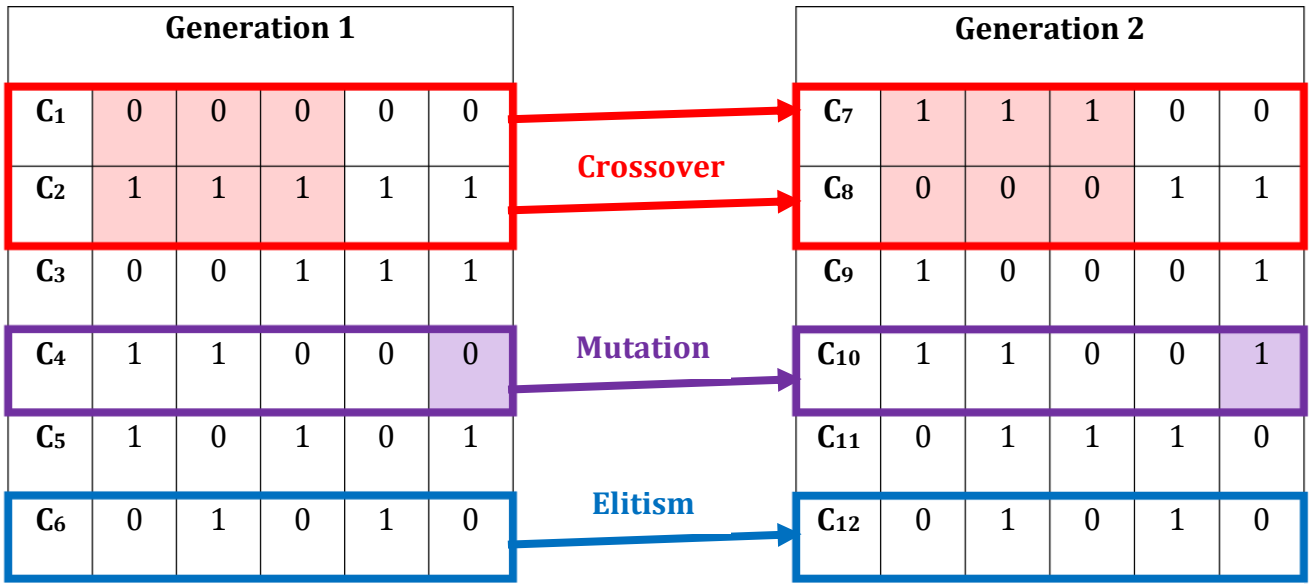


Figure 4.3. Summary of genetic processes that can occur between generations of a genetic algorithm. In red is crossover, where segments of chromosomes C_1 and C_2 are exchanged to produce two new solutions. In purple is shown mutation, where a single component of chromosome C_4 is altered to produce a new solution. In blue is the concept of elitism, where chromosome C_6 is carried on to the second generation, completely unchanged and treated as a new solution. These are simple examples of the concepts underlying the foundation of constructing a genetic algorithm.

In the context of our work, we have designed our genetic algorithm such that each solution, or chromosome, to our interface optimization problem can be represented by a specific combination of maps designated for each interface residue. Therefore, a population is represented by a specific group of these map combinations created at each generation. Cross-breeding our solutions would take place as a result of swapping a certain number of maps between two chromosomes and passing the two resulting offspring to the next generation. A mutation would occur when a specific map in a chromosome is altered to another possible map for that residue from our map library. In our model, elitism would move a top fraction of the best map combinations to a new generation, completely unaltered by crossover or mutation. With this GA incorporated, the overall workflow for our protein-protein interface optimization program will function according to the flowchart pictured in Figure 4.4. Our program is designed to begin with construction of libraries of ASCII-encoded

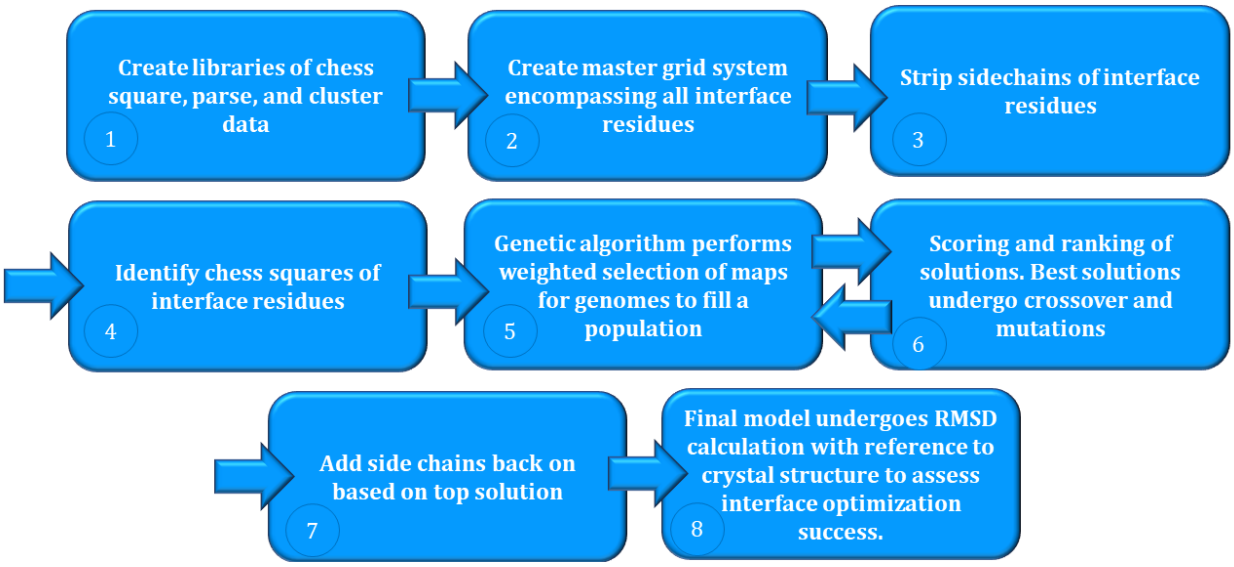


Figure 4.4. Workflow for our developing protein-protein interface optimization program. The program begins by creating libraries containing ASCII representations of addresses for our chess square, parse, and cluster data. We then construct a master grid system over all residues participating in interfacial interactions, where grid points are spaced 0.5 Å apart. The side chains from all interface residues are removed before we identify the chess squares into which they all fall. After all chess squares are known for each residue, our genetic algorithm performs a selection of map combinations to fill a population of a designated size, where the selection is weighted by relative populations of each cluster within each parse and chess square. All map combination solutions are scored and ranked before undergoing crossover and mutation processes. These two steps of the GA repeat for a number of times equal to a predetermined number of allowed generations. Finally, side chains are added back to interface residues based on the highest-scoring combination of maps when constructing the final model. The produced model undergoes an RMSD calculation with reference to the original crystal structure to assess the success of the docking.

libraries of our chess square, parse, and cluster data, immediately followed by construction of the previously described three-dimensional grid with 0.5 Å spacing (*vide supra*). All interface residues within the grid have their side chains removed and area assigned their appropriate chess square based on their backbone ϕ (phi) and ψ (psi) angles. Our genetic algorithm then performs a population-weighted selection of parse and cluster map combinations as a solution to the interface optimization problem with enough combinations to fill a population for that generation. All solutions are scored and ranked, according to our scoring system (*vide supra*). Certain solutions of that population undergo crossover and mutations, the products of which also undergo scoring and proceed to the next generation.

```
def crossover_option1():

A  #Generate Solutions
   Generate Population of Solutions
   Score and Rank All Solutions
   Select Solutions for Crossover and Mutation Operations

   #Select Parents and Breed
   #All possible parents are equal length arrays containing map selections
   #for each residue in the same order
   for parent1 in selected_solutions:
       random_integer1 = random number between 0 and len(parent1) #or len(parent2)
       random_integer2 = random number between random_integer1 and len(parent1)

       child1 = copy(parent1)
       child2 = copy(parent2)
       child1[random_integer1:random_integer2] = parent2[random_integer1:random_integer2]
       child2[random_integer1:random_integer2] = parent1[random_integer1:random_integer2]

       return child1 and child2

def crossover_option2():

B  #Generate Solutions
   Generate Population of Solutions
   Score and Rank All Solutions
   Select Solutions for Crossover and Mutation Operations

   #Select Parents and Breed
   #All possible parents are equal length arrays containing map selections
   #for each residue in the same order
   for parent1 in selected_solutions:
       Choose a random parent2
       Initialize child #same length as parent1 and parent2
       for residue in range(len(parent1)): #or parent2
           random_number = random number between 0 and 1
           if random_number < 0.5:
               child[residue] = parent1[residue]
           else:
               child[residue] = parent2[residue]

       return child
```

Figure 4.5. Pseudocode representing our genetic algorithm's crossover operation. A) Outlines crossover operation option 1, which is designed to swap maps in two segments of equal length and starting and ending at the same positions of two solutions. B) Outlines crossover operation option 2, which creates a child solution combination of maps of the same length as either of the two parents being bred. In this scenario, the residue map at any position in the solution array of maps is taken randomly from either parent at the same position. In this way, the crossover solution's genetic material is a product of a variety of different combination of the parents' genes.

```
def mutation():

    #Initialize Variables
    number_of_genomes = some_value
    mutation_rate = 0.05

    #Generate Solutions
    Generate Population of Solutions
    Score and Rank All Solutions
    Select Solutions for Crossover and Mutation Operations

    #Mutate Solution(s)
    for genome in range(number_of_genomes):
        random_number = random number between 0 and 1
        if random_number < mutation_rate:
            mutant_genome = genome.copy()
            random_residue = random number between 0 and len(genome)
            new_map =  population-weighted selection of new map for random_residue
            mutant_genome[random_residue] = new_map

    return mutant_genome
```

Figure 4.6. Pseudocode representing our genetic algorithm's mutation algorithm. This function depends on initialization of a certain predetermined number of genomes per population and mutation rate. After generation a population of solutions, ranking and scoring those solutions, and selecting the solutions that will undergo crossover and mutation operations, the chance of undergoing a mutation is applied to all solutions in the given population. A random residue position is selected, and the map at that residue position is swapped with another population weight-selected map, yielding a mutant form of the original model solution.

Examples of pseudocode for the crossover and mutation operations are visible in Figures 4.5 and 4.6, respectively. We decided on a flexible definition of "crossover," so we wrote two different functions for the crossover operation, which are both equally likely to happen. The population-filling and genetic process steps repeat for a number of times equal to a predetermined number of generations before the final, highest scoring model is constructed and compared to the original crystal structure via RMSD calculation. The true nature of this workflow is much more complicated and not quite as straightforward, as many of these steps are still undergoing thorough optimization, as will be described.

With all of this considered, it still is no easy task to develop the working features of a GA. It is important to consider the specific rates at which crossover and mutations occur, as

well as how many elite chromosomes are carried over from generation to generation. Where crossover is useful for swapping and sharing features/maps of high-scoring solutions among each other, mutations can be an effective way of perturbing the pool of existing high-scoring solutions and simulating the inclusion of features/maps that may not have been considered otherwise. Too much crossover can result in early convergence of solutions, and too much mutation can result in ineffective sampling of high-quality solutions. Many reports discussing the optimization of crossover and mutation rates have been published for this exact reason.[23-25] We have elected to begin constructing an adaptive genetic algorithm[26-28] that progressively alters the crossover and mutation rates, as well as the population size for each generation. Pseudocode representing the core adaptive component of this algorithm is given in Figure 4.7. This adaptive algorithm is designed to minutely increase the crossover rate and decrease the mutation rate when a generation produces a new high-scoring chromosome, in an attempt to slightly converge on a local minimum and prevent futile, overly broad sampling of random solutions. It also conversely decreases crossover and increases mutation when a generation *does not* produce a new high-scoring chromosome, in order to prevent sampling being trapped in local minima. Essentially, the algorithm will identify high-scoring combinations of maps and continue to explore any found local minima until such a point when the algorithm no longer identifies any high-scoring solutions in that minimum and it begins to explore minor variations in succeeding maps.

The next achievable goal of this project is to develop a working scoring system that correlates the scores of our output models with their RMSDs. Our theory is that higher-scoring combinations of our hydropathic maps will construct models of a protein-protein interface more similar to the original crystal structure on which the interface optimization

```
#Initialize Variables
number_of_genomes = some_value
crossover_rate = 0.1
mutation_rate = 0.01
rate_increase = 1.1
rate_decrease = 0.9
Overall_Best_Score = -99999999999.9

def adaptation():
    #Generate Solutions
    Generate Population of Solutions
    Score and Rank All Solutions
    Select Solutions for Crossover and Mutation Operations
    Perform Crossover and Mutation Operations
    Compile Highest Scoring Solutions
    Identify Generation Best Score

    #Adapt Crossover and Mutation Rates
    if Generation_Best_Score > Overall_Best_Score:
        Set Overall_Best_Score equal to Generation_Best_Score
        Multiply mutation_rate by rate_decrease
        Multiply number_of_crossover_rate by rate_increase
        Multiply number_of_genomes by rate_increase (?)

    elif Generation_Best_Score < Overall_Best_Score:
        Multiply mutation_rate by rate_increase
        Multiply number_of_crossover_rate by rate_decrease
        Multiply number_of_genomes by rate_decrease (?)

    return crossover_rate, mutation_rate, number_of_genomes
```

Figure 4.7. Pseudocode representing the major components of the adaptive genetic algorithm for our protein interface optimization tool. The code requires setting certain values for the number of genomes in a population, the number of crossover events that will occur during a generation, the mutation rate for solutions produced for the population and rates at which these features may increase or decrease. The overall best score is set to an extremely low and conquerable number by any model. A select group of high-scoring solutions must be chosen from the initial population. After all crossover events and mutations have occurred, a generation best score is identified, which is compared to the current overall best score. The number of crossover events, mutation rate, and number of genomes in a population are adjusted according to whether a new best overall scoring model was found.

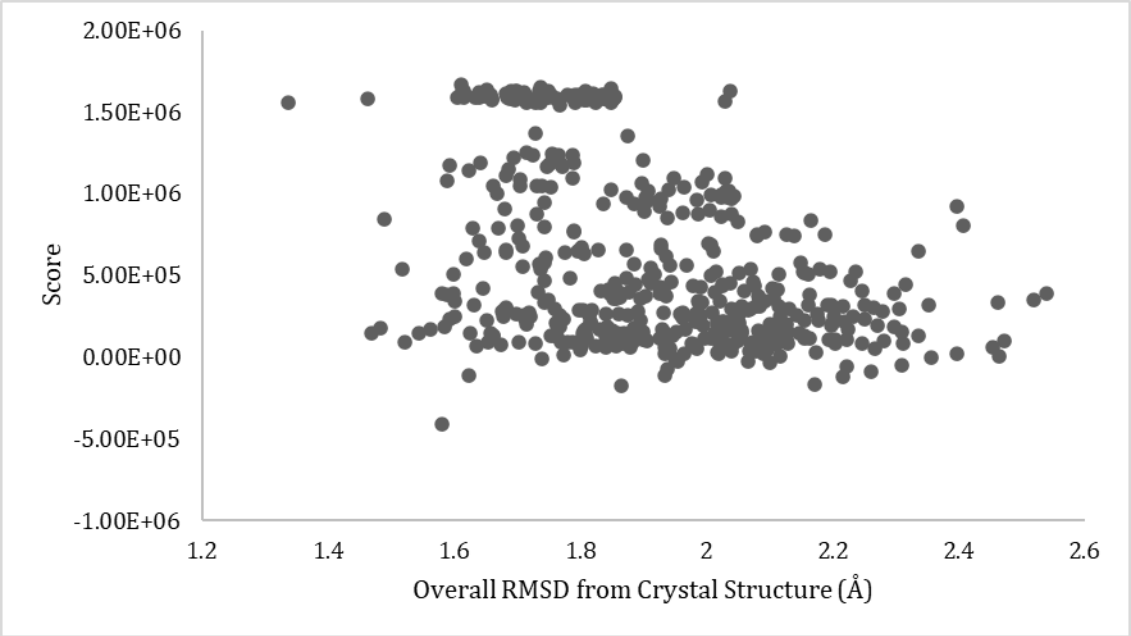


Figure 4.8. A plot of scores using our scoring function versus their overall RMSD values for 500 random models of the optimized protein-protein interface in the structure of PDB ID 2I25.[20] These models were not generated using the tools of the GA. A slight trend can be seen in this data, potentially illustrating a real relationship between our scoring function and RMSDs from template structures.

was conducted. We have attempted preliminary scoring to determine whether our current model successfully makes a correlation between our model scores and their RMSD values from the crystal structure. As an example of this preliminary scoring, see Figure 4.8. The plot in this figure shows scores of 500 random models generated without use of the genetic algorithm against their RMSDs from the crystal structure of PDB ID 2I25.[29] A slight downward trend is visible in this data, illustrating a potential relationship between our scores and their models' RMSDs, such that high scores correlate with a more similar structure to the crystallographic data. However, we still have much work to do to refine our model and achieve a better correlation with our data. Perhaps, one reason we do not have a high-quality correlation here is that these models were not constructed using the protocols of the GA. We sought to simply build models to ensure our scoring function could build the

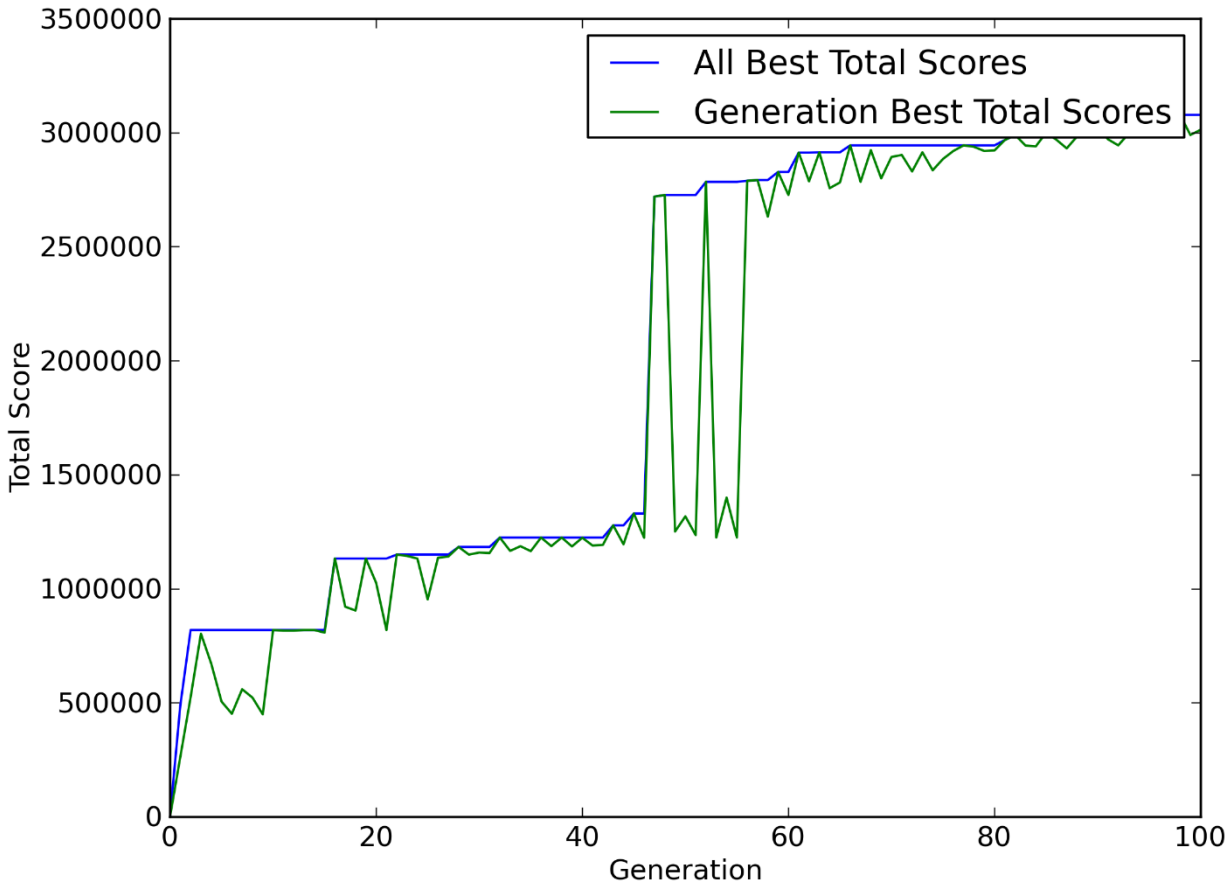


Figure 4.9. Example plot of best total scores of top-scoring models for each generation (green) and overall (blue). This plot was generated using the Python Matplotlib library as a product of interface residue position optimization of the two chains of PDB ID 1UZ3 over 100 generations. Plots such as this are being used to track performance of our genetic algorithm component of our interface optimization program. The many points at which the "Generation Best Total Scores" plot touches a plateaued "All Best Total Scores" plot indicates that it can take several generations for the crossover and mutation algorithms to find new, higher-scoring solutions. The steep jump in score possibly indicates a crucial mutation that appears to be a crucial component of a top-scoring model.

correlation we hoped for, but it is possible that allowing the GA to run and build higher-scoring models with, hopefully, even lower-RMSD values will yield a clearer relationship.

Another important step in the development of this tool is ensuring that the genetic algorithm is effectively sampling the solution space without converging early on local minima or wantonly sampling too many low-quality solutions due to the random nature of population-filling and the genetic processes. The adaptive components of the GA are

designed to mitigate these problems, but our model requires a method for assessing the success of these implementations. For this reason, optimization of our model includes charting the progressive change in score of our top-scoring model as the program runs. Currently, we have elected to simultaneously chart the score of the overall top-scoring model and the score of the top-scoring model for each generation as a way of tracking model improvement. An example of these plots is visible in Figure 4.9. This particular plot shows the progressive interface optimization of the two chains of PDB ID 1UZ3,[30] where the green plot indicates the score of the highest scoring solution produced each generation, while the blue plot indicates the change in the overall top-scoring solution across all generations. It can be seen from this plot that there are many periods during the progression of the GA that show very little change in the top score. The goal of producing these plots is to visually represent the performance of the GA that we might observe how changes to our crossover and mutation rates affect the rate of discovery of higher-scoring solutions. It is possible that a simple, but appropriate, metric for representing this rate of discovery could be the product of the slope of the "All Best Total Scores" plot and its R^2 as a way of considering the rate at which new, higher-scoring solutions are discovered and how often and long the top score plateaus, but this will require much further experimentation.

Current Challenges

We have made great progress toward developing our protein-protein docking tool, but we still have a number of challenges to face as we troubleshoot, optimize, decide best practices for how our program will operate. First and foremost among these challenges is finding more optimal crossover and mutation rates, as well as rates of change for the

adaptive component of the GA. Setting initial values for these rates can be somewhat arbitrary at first, so a great deal of experimentation is necessary to refine these values, which can be difficult while our program has extremely long runtimes. In our current model of the GA, the size of generation's population also scales with the mutation rate. Optimization of this parameter has also been shown to be a critical effector the design of an efficient GA.[26,31-33] Our current model scales population size with mutation rate because increasing these factors is designed to promote exploration of unexplored solution space. The crossover rate is conversely related, where it increases in response to decreasing population size and mutation rate because it is designed to explore solution space similar to discovered high-scoring solutions and close in on local minima. Nevertheless, the optimal rates for each of these factors is likely to be highly specific to our model and not easily determined.

Additionally, we have encountered some issues with many of our produced final models, wherein many side chains, particularly from the most flexible residues, like arginine and lysine, overlap with other side chains added back on to our final "docked" mode. An example of this problematic situation is visible in Figure 4.10. It is aggravating sometimes that these events occur in light of these models' high scores, but it is also expected, as the troublemaking residues are highly flexible ones. It must be stated our model does not include explicit punitive terms to actively discourage steric clashes, as our maps are designed to encode steric considerations based on the fact that the maps represent *spaces* where interactions take place. The interacting species, itself, occupies a void in space across from the residue on which the map is based. Essentially, we expect the most optimal solutions to benefit from overlapping maps and for sterically unfavorable/impossible, sub-optimal solutions to be out-competed where overlapping voids of space in maps are not rewarded.

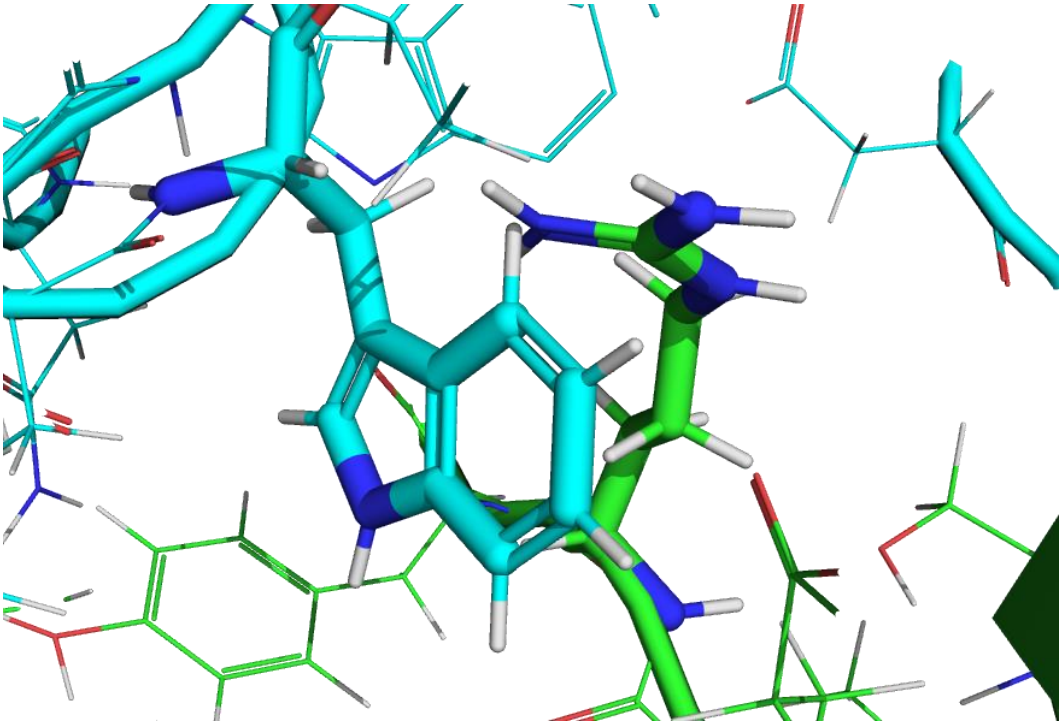


Figure 4.10. Interface optimization solution of PDB ID 2I25.[20] In this model created as a solution to the optimization of the two proteins in this crystal structure, two residues, a tryptophan (cyan) and an arginine (lime green) sterically clash. Our model currently does not penalize such unfavorable interactions.

In future work on this project, it may be wise to either consider including such terms, or disregard solutions that include such clashes, as they should not score high enough to be selected for crossover, mutation, or elitism anyway. It may also significantly reduce computational time to remove them from consideration before scoring them. This, too, will require a great deal of experimentation.

Lastly, possibly the largest and most complicated issue involves how certain map types are scored together. As previously mentioned, we are experimenting with scoring all map types with each other to determine whether a correlation can be seen between certain map type-type cross terms and our score output. We anticipate seeing little dependence on cross terms, such as positive hydrophobic-negative polar or negative hydrophobic-positive polar, but they require investigation, nonetheless. The more complicated aspect of this

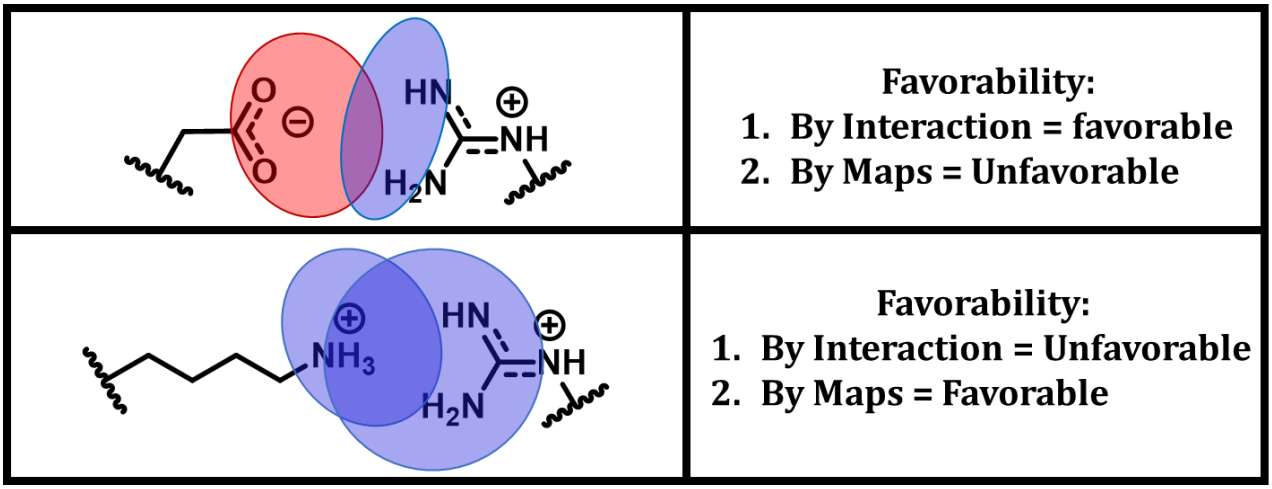


Figure 4.11. Example hypothetical, but realistic, scenarios that our residue optimization protocol may encounter. In the above two panels, an acidic residue interacts with an arginine, which is often seen in many real co-crystal structures. Normally, this is a highly favorable ionic interaction, but scoring overlapping positive and negative polar maps from these residues may deem this an unfavorable interaction. Likewise, in the bottom two panels, a lysine and arginine may cross paths in our algorithm. These residues have many maps with highly robust positive polar interactions that, if they interact, may identify this as a favorable interaction, when, in reality, it is not.

scoring system is that we must decide how to approach any type of polar interaction with another polar interaction. Hydrophobic-polar interactions are easy to penalize, as the hydropathic nature of the polar species is irrelevant, to an extent; simply put, hydrophobic does not interact favorably with polar of any nature. However, scoring polar interactions together can be tricky, as overlap of positive or negative polar maps with other polar maps does not explicitly define the hydropathic character of either polar species. The positivity or negativity of the interaction is completely relative and dependent on the acceptor or donor nature of the interacting partners. To illustrate this, Figure 4.11 shows how this issue may arise and the way it may be problematic for interface optimization. This figure represents two possible and realistic scenarios that may have already been encountered by our scoring algorithm. Situations like this can drastically affect the results of our scoring. It is not simply enough to consider the hydropathic valences for each residue indicated by our maps, which

are highly specific to the hydropathic nature of the residue in question, and how their interaction regions overlap. Therefore, we may need to incorporate some manner of post-processing consideration for hydropathic character beneath the interacting maps. Our maps are highly detailed in terms of the interaction character they capture, so it is important to ensure that the hydropathy of interacting species is considered.

Discussion

We have made significant progress in the development of a protein-protein docking tool, beginning with the construction of a protein interface optimization program. Our work in progress is designed to use the components of a dimeric protein complex, remove side chains from the interface residues, and determine the most favorable way to reconstruct them and their interactions with each other, hopefully reproducing the original crystal structure. Our plan is to utilize our vast hydropathic interaction map library, which exhaustively captures all unique interaction environments for each residue type. This data describes the interactions required to satiate the hydropathic character of a residue, giving us knowledge of how residues would prefer to interact with each other. These maps are information-rich and potentially useful for developing a number of different protein structure prediction tools in the long term.

Our approach to this protein-protein docking program has some strengths we wish to highlight, the first of which is essentially defined by our method, itself, that our maps simplify protein structure prediction problems down to a finite list of possible interactions a residue can form with its environment. Many approaches to protein structure prediction tools are designed to identify possible favorable interactions between residues and optimize

them as best as possible. With our unique method, we essentially have compiled a list of known manners of interaction between a residue and its environment, which are more detailed than simple knowledge of hydrogen bond donors interacting with acceptors. This could be likened to reading and knowing all possible answers to an exam prior to sitting for it. The problem is a matter of choosing the right combination of them, which is still easier than drafting them from scratch.

Another advantage of our approach is that our maps show, not only favorable interactions, but also unfavorable ones. While other approaches may focus on simply optimizing the most favorable interactions, such as hydrogen bonds and ionic interactions, our maps encode crucial information about disfavored ones in conjunction with favorable ones. Our hydropathic force field captured this information the same as any hydrogen bond, thus we believe them to be an integral part of protein structure and the way proteins fold and interact together.

Further, as described in Chapter 2, our method is capable of simulating variations in protonation state for ionizable residues. For residues, such as aspartic acid, glutamic acid, and histidine, we have calculated maps in a variety of different pH environments, which capture the hydropathic environments of residues in different protonation states. These maps, too, are incorporated into our map library. Because we effectively have twice the "normal" number of maps to choose from for these residues (i.e., maps calculated at high and low pHs), it should be noted that we have chosen first to use maps calculated at these residues' pH_{50}s (where half of all residues in our data set are protonated) to simplify map choice, but in the final adaptation of our program, we plan to make high pH (deprotonated)

and low pH (protonated) residue maps available for docking. The advantage of implementing this data is that varying protonation state may completely alter the favorability of a map-map interaction based on the change from a hydrogen bond acceptor to a donor for a single residue. On this same subject, our system, which is still based on the HINT force field, is highly amenable to insertion of water molecules at key positions that would optimize certain unfavorable interactions into favorable ones. What we wish to emphasize here, as we have done elsewhere, is that our method is highly adaptable and requires a significant amount of experimentation to modify and optimize certain features of the program.

Finally, there are certain niche interactions that are encoded by our maps that may by important to consider for a problem, such as protein-protein docking. These particular interactions include π-π, π-cation, and possible formation of cystine bonds. Through the chapters in 3DIH we have published, where we have described the capture of these interactions in our maps, we have compiled this data as part of our map library to be included in docking solutions, which, as far as we know, is unique to our method. Though these interactions may be less common than traditional hydrogen bonding, they are no less important to consider when conducting a protein-protein docking experiment.

In spite of the advantages of our method, we have a number of challenges to overcome as we continue to optimize parts of our algorithm. Among them are the progressive refinement of our crossover and mutation rates, population size, and adaptive parameters. Our goal with this endeavor is to prevent early convergence and curb sampling of low-quality, low-scoring solutions. This will be especially difficult because we also currently face extremely long run times for our program for runs using more generations and larger

populations, which are necessary to effectively probe the success of these features. We are also finding that many of output models place some residues on top of others, which we are currently examining as an issue with maps of our more flexible residues. We have also proposed conducting energy minimizations post-model construction, which we are still investigating. Another possible solution is to penalize sterically improbable or impossible map combinations, which we will have to explore later. Lastly, must resolve how certain polar interactions should be considered. Overlap of positive and/or negative polar maps is problematic because the favorability of the interaction underneath our maps is completely relative to the hydropathy of the interacting species, where the maps alone do not describe the favorability of the interaction. One possible approach is to write post-processing code to identify the interaction partners and whether it is a truly favorable or unfavorable interaction. The model we have constructed certainly does face its fair share of challenges, but we have solutions in mind that we are confident will progress us toward a functioning protein-protein docking tool.

Conclusions

Protein structure prediction has made many leaps and bounds in the past decade of work, but researchers have work cut out for them, still. Much work has yet to be done, but our group has made many strides in the direction of multimeric interaction prediction in the form of this protein-protein interface optimization tool discussed here. Our goal has been to combine data-rich hydropathic interaction maps unique to our lab and apply them to a protein-protein docking tool, knowing these maps are encoded with information concerning the interactions residues will make with each other. We are confident that, in the future, we

can build a fully functioning program that will be freely available for use by other investigators.

References

1. Vakser, I. A. Protein-Protein Docking: From Interaction to Interactome. *Biophys. J.* **2014**, *107*, 1785-1793.
2. Goncearenco, A.; Li, M.; Simonetti, F. L.; Shoemaker, B. A.; Panchenko, A. R. Exploring Protein-Protein Interactions as Drug Targets for Anti-Cancer Therapy with In-Silico Workflows. *Methods Mol. Biol.* **2017**, *1647*, 221-236.
3. Guntas, G.; Purbeck, C.; Kuhlman, B. Engineering a Protein-Protein Interface Using a Computationally Designed Library. *Proc. Natl. Acad. Sci. U. S. A.* **2010**, *107*, 19296-19301.
4. Elcock, A. H.; Sept, D.; McCammon, J. A. Computer Simulation of Protein-Protein Complexes. *J. Phys. Chem. B* **2001**, *105*, 1504-1518.
5. Samsonov, S. A.; Teyra, J.; Anders, G.; Pisabarro, M. T. Analysis of the Impact of Solvent on Contacts Prediction in Proteins. *BMC Struct. Biol.* **2009**, *9*, 22.
6. Joachimiak, L. A.; Kortemme, T.; Stoddard, B. L.; Baker, D. Computational Design of a New Hydrogen Bond Network and At Least a 300-Fold Specificity Switch at a Protein-Protein Interface. *J. Mol. Biol.* **2006**, *361*, 195-208.
7. Keskin, O.; Nussinov, R. Similar Binding Sites and Different Partners: Implications to Shared Proteins in Cellular Pathways. *Structure*, **2007**, *15*, 341-354.

8. Ahmed, M. H.; Spyrakis, F.; Cozzini, P.; Tripathi, P. K.; Mozzarelli, A.; Scarsdale, J. N.; Safo, M. A.; Kellogg, G. E. Bound Water at Protein-Protein Interfaces: Partners, Roles and Hydrophobic Bubbles as a Conserved Motif. *PLoS One* **2011**, *6*, e24712.
9. Ahmed, M. H.; Habtemariam, M.; Safo, M. K.; Scarsdale, J. N.; Spyrakis, F.; Cozzini, P.; Mozzarelli, A.; Kellogg, G. E. Unintended Consequences? Water Molecules at Biological and Crystallographic Protein-Protein Interfaces. *Comp. Biol. Chem.* **2013**, *47*, 126-141.
10. Gray, J. J.; Moughon, S.; Wang, C.; Schueler-Furman, O.; Kuhlman, B.; Rohl, C. A.; Baker, D. Protein-Protein Docking with Simultaneous Optimization of Rigid-Body Displacement and Side-Chain Conformations. *J. Mol. Biol.* **2003**, 281-299.
11. Yan, Y.; Tao, H.; He, J.; Huang, S. The HDOCK Server for Integrated Protein-Protein Docking. *Nat. Protoc.* **2020**, *15*, 1829-1852.
12. Kozakov, D.; Hall, D. R.; Xia, B.; Porter, K. A.; Padhorny, D.; Yueh, C.; Beglov, D.; Vajda, S. The ClusPro Web Server for Protein-Protein Docking. *Nat. Protoc.* **2017**, *12*, 255-278.
13. Parikh, H. I.; Kellogg, G. E. Intuitive, But Not Simple: Including Explicit Water Molecules in Protein-Protein Docking Simulations Improves Model Quality. *Proteins* **2013**, *82*, 916-932.
14. Herrington, N. B.; Kellogg, G. E. 3D Interaction Homology: Computational Titration of Aspartic Acid, Glutamic Acid, and Histidine Can Create pH-Tunable Hydropathic Environment Maps. *Front. Mol. Biosci.* **2021**, *8*, 773385.
15. Ahmed, M. H.; Koparde, V. N.; Safo, M. K.; Neel Scarsdale, J.; Kellogg, G. E. 3D Interaction Homology: The Structurally Known Rotamers of Tyrosine Derive from a Surprisingly

Limited Set of Information-Rich Hydropathic Interaction Environments Described by Maps. *Proteins* **2015**, *83*, 1118–1136. doi:10.1002/prot.24813

16. Ahmed, M. H.; Catalano, C.; Portillo, S. C.; Safo, M. K.; Neel Scarsdale, J.; Kellogg, G. E. 3D Interaction Homology: The Hydropathic Interaction Environments of Even Alanine Are Diverse and Provide Novel Structural Insight. *J. Struct. Biol.* **2019**, *207*, 183–198. doi:10.1016/j.jsb.2019.05.007
17. AL Mughram, M. H.; Catalano, C.; Bowry, J. P.; Safo, M. K.; Scarsdale, J. N.; Kellogg, G. E. 3D Interaction Homology: Hydropathic Analyses of the "π-Cation" and "π-π" Interaction Motifs in Phenylalanine, Tyrosine, and Tryptophan Residues. *J. Chem. Inf. Model*. **2021**, *61*, 2937–2956. doi:10.1021/acs.jcim.1c00235
18. Catalano, C.; AL Mughram, M. H.; Guo, Y.; Kellogg, G. E. 3D Interaction Homology: Hydropathic Interaction Environments of Serine and Cysteine Are Strikingly Different and Their Roles Adapt in Membrane Proteins. *Curr. Res. Struct. Biol.* **2021**, *3*, 239–256. doi:10.1016/j.crstbi.2021.09.002
19. Kellogg, G. E.; Semus, S. F.; Abraham, D. J. HINT: A New Method of Empirical Hydrophobic Field Calculation for CoMFA. *J. Comput. Aided Mol. Des*. **1991**, *5*, 545-552.
20. Kellogg, G. E.; Abraham, D. J. Hydrophobicity: Is LogP(o/w) More Than the Sum of its Parts? *Eur. J. Med. Chem*. **2000**, *35*, 651-661.
21. Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, B. E. The Protein Data Bank. *Nucleic Acids Res*. **2000**, *28*, 235-242.
22. Shapovalov, M. V.; Dunbrack, R. L., Jr. A Smoothed Backbone-Dependent Rotamer Library for Proteins Derived from Adaptive Kernel Density Estimates and Regressions. *Structure* **2011**, *19*, 844–858. doi:10.1016/j.str.2011.03.019

23. Hassanat, A.; Almohammadi, K.; Alkafaween, E.; Abunawas, E.; Hammouri, A.; Surya Prasath, V. B. Choosing Mutation and Crossover Ratios for Genetic Algorithms – A Review with a New Dynamic Approach. *Information* **2019**, *10*, 390.
24. Patil, V. P.; Pawar, D. D. The Optimal Crossover or Mutation Rates in Genetic Algorithm: A Review. *Int. J. Appl. Eng. Technol.* **2015**, *5*, 38-41.
25. Greenwell, R. N.; Angus, J. E.; Finck, M. Optimal Mutation Probability for Genetic Algorithms. *Math. Comput. Model.* **1995**, *21*, 1-11.
26. Ye, Z.; Li, Z.; Xie, M. Some Improvements on Adaptive Genetic Algorithms for Reliability-Related Applications. *Reliab. Eng. Syst. Saf.* **2010**, *95*, 120-126.
27. Yun, Y.; Gen, M. Performance Analysis of Adaptive Genetic Algorithms with Fuzzy Logic and Heuristics. *Fuzzy Optim. Decis. Mak.* **2003**, *2*, 161-175.
28. Bingul, Z. Adaptive Genetic Algorithms Applied to Dynamic Multiobjective Problems. *Appl. Soft Comput.* **2007**, *7*, 791-799.
29. Chavali, G. B.; Ekblad, C. M. S.; Basu, B. P.; Brissett, N. C.; Veprintsev, D.; Hughes-Davies, L.; Kouzarides, T.; Itzhaki, L. S.; Doherty, A. J. Crystal Structure of the ENT Domain of Human EMSY. *J. Mol. Biol.* **2005**, *350*, 964-973.
30. Stanfield, R. L.; Dooley, H.; Verdino, P.; Flajnik, M. F.; Wilson, I. A. Maturation of Shark Single-Domain (IgNAR) Antibodies: Evidence for Induced-Fit Binding. *J. Mol. Biol.* **2007**, *367*, 358-372.
31. Rajakumar, B. R.; George, A. APOGA: An Adaptive Population Pool Size Based Genetic Algorithm. *4*, **2013**, 288-296.

32. Diaz-Gomez, P. A.; Hougen, D. F. Initial Population for Genetic Algorithms: A Metric Approach. *In Proceedings of the International Conference on Genetic and Evolutionary Methods, Las Vegas, USA*, **2007**, 25-28.
33. Koljonen, J.; Alander, J. T. Effects of Population Size and Relative Elitism on Optimization Speed and Reliability of Genetic Algorithms. *In Proceedings of the 12th Finnish Artificial Intelligence Conference STeP, Finland*, **2006**, 26-27.

Chapter 5: Conclusions

It has been stated numerous times in this thesis already, but it should be re-emphasized that the advances in computational chemistry, structural biology, and molecular modeling have put knowledge of protein structure at the forefront of modern drug discovery. The studies enclosed in this thesis demonstrate that and offer a variety of perspectives into the utility of publicly available structural data. They also further showcase the importance of collecting and refining this structural data for employing it to other structural studies and drug discovery campaigns. Our lab's primary focus rests on developing new protein structure prediction programs to provide guidance and new tools to drive new efforts in these areas.

The first study presented here is an update to our ongoing '3D Interaction Homology' project, where we attempted to study the hydropathic interaction environments of ionizable residues aspartic acid, glutamic acid, and histidine, determine the environments that contribute to stabilization of their different ionization states, and use this information to construct hydropathic maps to be used later in protein structure prediction tools undergoing development. We designed an algorithm using our in-house HINT force field to simulate the free energy change of altering residues' ionization states at different pHs. Further, we learned a great deal about the unique roles of each residue type in protein structures and how we may exploit these features in protein structure prediction tools. To this end, we calculated "low pH" and "high pH" maps for these residues to be incorporated into tools of our design. Our work on this project is important, mostly because ionization state optimization is a largely unaddressed in modern protein structure prediction tools, including

(perhaps, especially) that of AlphaFold. We hope that our work promotes recognition of the issue we have addressed and drives efforts to improve our consideration of a crucial feature of protein structure modeling.

Secondly, this thesis shares a "traditional" computational drug discovery story that identifies highly potent inhibitors of eIF4A1, a validated cancer target. We were interested in the discovery of novel inhibitors against a target with largely untapped potential for therapeutic value. Pharmacophoric virtual screening of a natural product inhibitor co-crystallized with a eIF4A1:RNA complex rewarded us with an extremely potent hit with an IC_{50} of approximately 1 nM in our luciferase-based readout assay. As remarkable as this was, purchase of analogues of this hit identified even more potent compounds with minor structural variations. We attempted numerous modeling studies to determine the binding mode of this series of compounds, which resulted in construction of a reasonable model that may explain the drastic increase in activity from the initial hit to our lead compound. Further, our model will be the basis for design of new, hopefully more potent and drug-like compounds as this drug discovery campaign proceeds. This work, in particular, testifies to the sheer power of computational modeling resources for predicting active drug-like therapeutics. The fact that *in silico* tools can reliably provide a route to discovery of new medicines is remarkable.

The final study described here explores our first attempt to employ use of our hydropathic environment map data in development of a protein structure prediction tool. This tool, still undergoing optimization, represents the first step toward a protein-protein docking tool that will predict how two proteins will interact together on a residue scale. This

precursor step is designed as a protein-protein interface optimization program that will use the backbone structures of two already co-crystallized proteins and determine the most favorable interactions for their interface residues and reproduce the original crystal structure. Both the genetic algorithm and scoring function at the core of this program are still undergoing development and optimization. We are experimenting with variations in crossover and mutation rates of the genetic algorithm and exploring combinations and scaling of hydropathic map interaction terms, but our initial attempts at modeling protein-protein interfaces suggest that our scoring function can be correlated with RMSD from the original crystal structure. This project has made leaps and bounds, but still requires significant work and time to construct a working tool. However, the potential impact of this tool will further many drug discovery efforts, as disrupting protein-protein interactions is still being explored as a route for new therapies. It will also represent a major milestone for our ongoing protein structure prediction work as the first implementation of our hydropathic map data.

Much work has been done and discussed to produce this thesis. It is a monument to the diligent work and creative thinking of many scientists to further projects that may eventually translate into very real medicines for patients in need. To progress this far, vast amounts of structural data were required, but this demonstrates the ultimate value of and need for accurate protein structure modeling in modern drug discovery. The future of novel drug discovery will be significantly impacted by our advances in computational chemistry and modeling, and this thesis is testament to that notion.

Vita

Noah Benjamin Herrington was born in Richmond, Virginia to parents Janice Silver and Alfred Herrington III. He was raised in Chesterfield, Virginia before graduating as Valedictorian from Meadowbrook High School and attending Randolph-Macon College for his undergraduate studies, where he received his Bachelor of Science in Chemistry and Spanish in 2018. In the Fall of that year, he later joined the Department of Medicinal Chemistry at Virginia Commonwealth University to pursue his Doctor of Philosophy under the guidance of Dr. Glen E. Kellogg, Ph.D.

Publications (†Contributed Equally)

1. Kayastha, F.†; Herrington, N. B.†; Kapadia, B. †; Roychowdhury, A.; Nanaji, N.; Kellogg, G. E.; Gartenhaus, R. B. A Novel eIF4A1 Inhibitor with Anti-Tumor Activity in Diffuse Large B-Cell Lymphoma. *Molecular Medicine.* submitted.
2. AL Mughram, M. H.†; Herrington, N. B.†; Catalano, C.; Kellogg, G. E. Systematized Analysis of Secondary Structure Dependence of Key Structural Features of Residues in Soluble and Membrane-Bound Proteins. *J. Struct. Biol.: X* **2021**, *5*, 100055 (https://doi.org/10.1016/j.yjsbx.2021.100055).
3. Herrington, N. B.; Kellogg, G. E. 3D Interaction Homology: Computational Titration of Aspartic Acid, Glutamic Acid and Histidine Can Create pH-Tunable Hydropathic Environment Maps. *Front. Mol. Biosci.*, **2021**, *8*, 773385 (https://doi.org/10.3389/fmolb.2021.773385).
4. Perry, C. K.; Casey, A. B.; Felsing, D. E.; Vermula, R.; Zaka, M.; Herrington, N. B.; Cui, M.; Kellogg, G. E.; Canal, C. E.; Booth, R. G. Synthesis of Novel 5-Substituted-2-Tetralin Analogs: 5-HT1A and 5-HT7 G Protein-Coupled Receptor Affinity, 3D-QSAR and Molecular Modeling. *Bioorg. Med. Chem.* **2020**, *28*, 115262 (https://doi.org/10.1016/j.bmc.2019.115262).

Honors

1. School of Pharmacy Research & Career Day Poster Presentation Runner-Up, Virginia Commonwealth University, March 2021.

2. Lester W. Morris, Jr. Scholarship, November 2020.

3. Graduate Poster Competition Winner, *The Protein Society 35th Anniversary Symposium*, July 2021.

4. Graduate School Travel Grant, Virginia Commonwealth University, March 2022.

5. Rector & Rorrer Travel Award, School of Pharmacy, Virginia Commonwealth University, March 2022.

“The problem with troubleshooting is that trouble shoots back.” – Ben Aaronovitch