Computer-aided drug design: developing and applying simulation-based tools to identify small-molecule ligands that inhibit proteins

by

Erich Hellemann Holguín

B. Sc. Chemistry, Universidad Nacional Autónoma de México, 2011

M. Sc. Chemistry, Carnegie Mellon University, 2016

Submitted to the Graduate Faculty of the

Dietrich School of Arts and Sciences in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2022

UNIVERSITY OF PITTSBURGH

DIETRICH SCHOOL OF ARTS AND SCIENCES

This dissertation was presented

by

Erich Hellemann Holguín

It was defended on

September 21, 2022

and approved by

Maria Kurnikova, Professor, Department of Chemistry

Lillian T. Chong, Associate Professor, Department of Chemistry

David R. Koes, Associate Professor, Department of Computational and Systems Biology

Thesis Advisor: Jacob D. Durrant, Associate Professor, Department of Biological Sciences

Computer-aided drug design: developing and applying simulation-based tools to identify small-molecule ligands that inhibit proteins

Erich Hellemann Holguín, PhD

University of Pittsburgh, 2022

In this dissertation, I discuss how computational methods can help in drug discovery, from developing a new tool that allows obtaining ensembles of protein conformations to using established computational tools for elucidating the mechanism of inhibition.

Sub-Pocket Explorer (SubPEx) is a tool I wrote that leverages weighted ensemble to accelerate the sampling of protein pocket conformations. I demonstrated that SubPEx is faster and protein pocket conformations are more diverse than those obtained by vanilla molecular dynamics (MD) simulations. I applied the SubPEx algorithm to three relevant proteins for human health: heat shock protein 90, neuraminidase, and hexokinase II. With these proteins, I showed how SubPEx could be applied to small rigid proteins, proteins with a flexible pocket, and proteins that undergo extensive domain rearrangements.

I show how a combination of experimental and computational work can help find a new resistance mechanism to the known inhibitor 2-deoxy-glucose (2DG). I described how collaborators found a new mutation in hexokinase II and how with MD simulations, I proposed a mechanism by which this mutation could confer resistance to 2DG.

Finally, I show the importance of undergraduate research and what we can achieve with the help of undergraduates. In one of the two projects I did with undergraduates, we applied an established computational protocol to recommend small molecule binders to a protein involved in cancer. The last project I described in the dissertation is how we used MD simulations to discover

the allosteric mechanism by which a small molecule inhibits TEM-1, a protein involved in multidrug resistance.

Table of Contents

List of Tables

List of Figures

Preface

"Por mi raza hablará el espíritu" – José Vasconcelos.

This dissertation is dedicated to the love of my life and the lovable tiny monster we created. I could not hope for a better partner in life than both of you, Paulette and baby F. This was a tortuous and lengthy endeavor that we both had to go through, but in the end, it will be worth it. We survived the good and the bad we encountered on this path, which only made us stronger and more resilient. We laughed, we cried, we spent countless hours just enjoying the company we had, and then a little hurricane came into our lives and made everything more complicated, but it also brought a really fulfilling new stage in our lives. Paulette, I love you. Thank you for believing in me even though I sometimes doubted myself.

I also want to thank my Ph.D. advisor, Dr. Jacob Durrant, and the people in the laboratory. Jacob created an environment where we could explore and learn, and above all, we were respected. Under your supervision, I became the scientist I am today, and I know we will still be working together for a long time. To the laboratory mates, I want to thank you for our discussions, for bouncing ideas off of each other, and for letting me say all the bad jokes I told. You are free of my bad jokes! To the undergraduates I mentored, thank you for your hard work! I hope you are as proud of me as I am of you! I know the three of you will do great things, and I can't wait to see it.

To my family, I am grateful for your upbringing and how you set me up for success; I think getting a Ph.D. is what some people call success. Thank you for always loving me, even when I know I can be a little bit stubborn and hard to deal with. ¡Los quiero con el alma!

And finally, to my friends, I want to thank you for the fun times, for being there for me, and for the times we went to sleep late playing a board game or some other dumb thing. I love you all.

1.0 Introduction

1.1 Dissertation outline

This dissertation begins with an introduction to computer-aided drug design that describes techniques used in the following chapters. After the introduction, three chapters detail projects I contributed to.

In chapter two, I describe the development of Sub-Pocket EXplorer (SubPEx), a tool that uses weighted ensemble simulations to sample protein pocket conformations efficiently. I explain how I developed the progress coordinate used in SubPEx, and how I cluster the simulations to obtain representative conformations. I finish by testing SubPEx on neuraminidase and hexokinase II.

In chapter three, I describe a collaboration between experimentalists and computationalists. We found a new mutation in *Saccharomyces cerevisiae hexokinase II that confers* 2-deoxy-glucose resistance. I used molecular dynamics in this work to understand how this mutation confers resistance to 2DG.

Chapter four describes projects where I mentored undergraduate students. I start with a brief introduction of how research in undergraduate studies is advantageous to the student and the field. The first project involved searching for ligands that could inhibit the activity of SMUG1, a protein responsible for repairing DNA damage. We used homology modeling to obtain a series of protein conformations for use in ensemble docking. The second project used molecular dynamics to understand the mechanism by which an allosteric ligand inhibits TEM-1, a known antibiotic-resistance-conferring protein.

Finally, in chapter five, I finish with some concluding remarks and future directions.

1.2 Computer-aided drug design

Drug discovery is a time-consuming and expensive endeavor. Studies propose costs ranging from $161 million to a staggering $4.54 billion to bring a new drug to the market, with development times between 10 and 15 years.[1,2] Given these high costs and the decline of new molecular entities (NMEs) being approved by the FDA, the pharmaceutical industry and research laboratories have searched for ways to lessen the costs and increase success rates (the FDA approves only about 13% of drugs entering clinical trials).[3]

One approach involves a combination of experimental and computational methods.[4] Computer-aided drug design or discovery (CADD) is a collection of computational techniques that have permeated almost all steps in the drug discovery pipeline. CADD is a rational approach to drug development. It has become essential in early-stage lead discovery since it aims to minimize failures cost-effectively. One of the early successes of CADD is the development of dorzolamide, a carbonic anhydrase inhibitor. Here, scientists used first-principle calculations to understand the differences in inhibition between two enantiomers.[5,6]

CADD can be classified into two approaches: structure-based drug design (SBDD) and ligand-based drug design (LBDD).

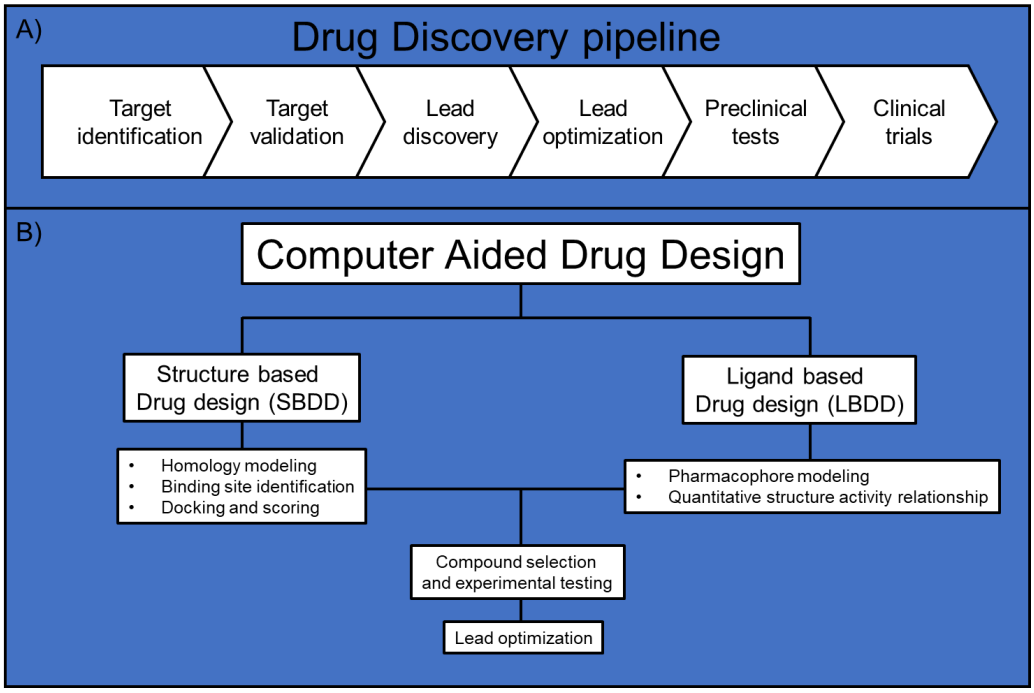


Figure 1. Drug discovery pipeline and Computer Aided Drug Design. Image adapted from Sumudu P. Leelananda et al.[7] A) Steps in the drug discovery pipeline. B) Computer-aided drug design can be divided into two branches: ligand and structure-based drug design. Presented are some of the techniques used in CADD. Once one or more compounds have been found, experimental validation is needed.

The focus of this thesis is going to be on SBDD. SBDD is a process of rationally designing new drugs, using many tools/methods to accomplish that goal. The main requirement in SBDD is to have an atomic resolution structure of the protein target; when the protein target is not known, or a structure cannot be obtained, LBDD is used.8 SBDD leverages the macromolecule's three-dimensional structure to predict interactions with prospective ligands in hopes of finding a small molecule that will progress through clinical trials, to end up in the market as therapeutics finally. In the following sections, I will briefly introduce some of the computational tools used in SBDD.

1.2.1 Homology modeling

Usually, the goal in drug discovery is to use protein structures determined by X-ray crystallography, nuclear magnetic resonance (NMR), or cryo-electron microscopy (cryo-EM). It is not always possible to use experimentally derived protein structures; when they are unavailable, one can predict the three-dimensional structure of a protein using computational methods. Homology modeling, also known as comparative modeling, is a computational technique that takes a protein amino-acid sequence to predict that three-dimensional structure. It takes advantage of the fact that protein structure is more conserved during evolution than is the sequence, so sequence-related proteins generally share similar structures.[8]

The homology-modeling algorithm starts by identifying template proteins with known three-dimensional structures. A multi-template approach can be used when no single template covers the whole protein. When there is a significant sequence identity between the protein and a template (30% or higher), an approximate structure can likely be obtained. Once suitable templates have been found, sequence alignment between the protein and its templates is done. This step is crucial in model accuracy and uses pairwise sequence alignment. Then, the algorithm uses the templates to generate an initial model for the protein, which is later refined. Refinement usually includes a minimization step that uses molecular mechanics. The last step is model validation or evaluation, which is when the model accuracy is determined.[9,10]

Lately, machine learning (ML) approaches have been used to generate models. The catalyzer for ML's use in homology modeling is the advent of the "big data" era.[11,12] Specifically, convolutional neural networks (CNN), an algorithm used in image analysis, have been used to predict protein structures, as exemplified by AlphaFold and RaptorX.[13,14] According to the authors,

AlphaFold is the first tool that predicts protein structures "to near experimental accuracy in a majority of cases."[14]

As a rule, a model is considered suitable for structure-based drug discovery if the sequence identity is above 50%; these models are considered to be of sufficient quality for SBDD. Models generated with templates that share 25-50% sequence identity can be used to assess how druggable a protein target is. Models with less than 25% sequence identity are only useful for directing mutagenesis experiments or protein function assignment.[10,15]

There are plenty of examples of how homology modeling has helped in the drug discovery process, but there are still challenges to overcome. The first and foremost challenge is how homology modeling struggles to predict the structure of proteins whose family structure has not been determined and the structure of intrinsically disordered proteins or intrinsically disordered protein regions.[12]

1.2.2 Molecular docking

High-throughput screening (HTS), introduced in the early 90s, is an experimental technique used to accelerate lead discovery.[16,17] In HTS, an extensive library of small molecules is tested against a target, often using an *in vitro* essay (*e.g.,* fluorescence) to observe a response and find drug candidates. The objective of HTS is to test thousands of compounds per day to find a small-molecule that modulates the activity of a specific target, be it a protein, a pathway, or a cellular event.[18] For HTS to be able to test the massive amount of compounds per day, it needs to use automation of sample handling, assay processing, and response readouts. HTS is expensive because you must have the equipment, reagents, and compound libraries, and you have to maintain

everything. This problem is exacerbated by the development of combinatorial chemistry, which has exponentially increased the size of many available chemical libraries.[19]

To reduce the number of compounds that need to be experimentally tested, the research community has developed computational methods to screen molecules *in silico*. Virtual screening uses experimental or physical principles to predict compounds that will be active, reducing the number of molecules that need to be tested experimentally.[20]

The most used SBDD technique for virtual screening is molecular docking. The goal of docking is to predict the binding affinity between a protein and a ligand. Docking algorithms seek to answer two questions. First, what is the ideal pose (spatial arrangement) of the ligand in the protein pocket? Second, how strong is this interaction (scoring)? The algorithm must be fast and accurate if one hopes to apply docking in virtual screening.

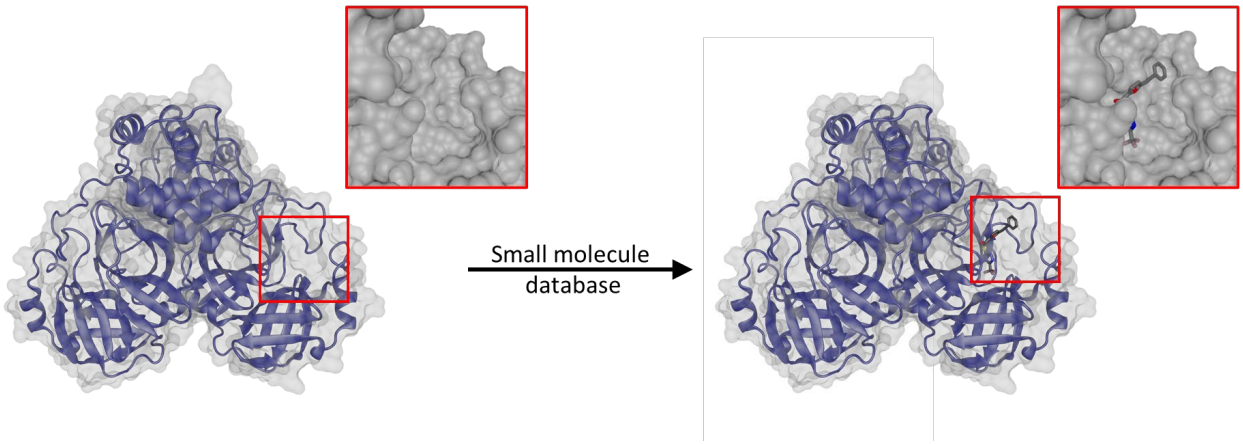


Figure 2. Simplified depiction of how molecular docking can find a drug lead. First, a protein pocket has to be identified with the goal that a small molecule will bind to it and elicit a biological response. Once the pocket has been identified, one tries to predict how small molecules will bind and the strength of this binding.

In practice, it is computationally expensive to sample all the poses a ligand can assume within the pocket. Consequently, most searching algorithms consider the protein as a rigid body and only sample a fraction of the ligand's available conformational space.[21] One popular searching approach is the Monte Carlo algorithm.[20–22] This method uses simulated annealing and random

moves to sample across the energy landscape. To accept or reject the proposed move, it uses the Metropolis criteria.

Once a single or a set of ligand poses has been found by the searching algorithm, the docking program predicts the strength of the protein-ligand interaction. Scoring functions, which can be physics-based, knowledge-based, or empirical, are used to this effect.[23] These functions serve two purposes. First, they are used to find the best pose for a ligand. Second, after docking a library of small molecules, they help differentiate between possible and poor binders.

However, proteins are inherently flexible; they "wiggle and jiggle,"* which allows them to adopt different conformations. In some cases, substantial changes can be seen in the protein binding pocket. Molecular docking typically ignores this conformational richness. However, by docking a small molecule database to an ensemble of protein conformations, we incorporate some of this conformational richness into the virtual screen. This technique is called ensemble docking and has been shown to improve predictions of binding compared to single conformation docking.[24,25] It is preferable to get the protein conformations through experimental means (X-ray or NMR), but that is not always possible; for such cases, one can use computational methods, like molecular dynamics simulations.

* "If we were to name the most powerful assumption of all, which leads one on and on in an attempt to understand life, it is that all things are made of atoms, and that everything that living things do can be understood in terms of the jigglings and wigglings of atoms." - Richard P. Feynman

1.2.3 Molecular dynamics

Proteins perform many of the biochemical functions that cells require for survival. This function is encoded by the protein structure and its dynamics. X-ray crystallography, NMR, and cryo-EM are excellent experimental techniques for obtaining protein structures, but these "pictures" provide only a static representation of the protein. We can use computational simulations to get an atomistic "movie" that captures the jigglings and wigglings, which often provide meaningful information about their function. Quantum mechanics calculations are impossibly expensive for proteins, so we need to simplify the biomolecular system we want to study. We can use Newtonian physics to approximate protein motions. To do this, we need to assume that the atoms' position is represented by their nuclear positions (Born-Oppenheimer approximation).[26,27] Other approximations are how we represent atoms and bonds, and they are treated as solid spheres and springs, respectively.

Molecular dynamics simulations consist of three steps that we repeat over and over again. First, we calculate a set of approximate forces for each atom using a semi-empirical force filed. The force filed approximates the system by using a potential energy function that models bonding, bond angles, bond torsions, electrostatics, and van der Waals interactions (see equation below). We next move the atoms according to these forces and advance the simulation by one or two quadrillionths of a second (1 or 2 fs). We repeat these three steps millions or billions of times.

$$v(r^N) = \sum_{bonds} \frac{k_r}{2}(l_i - l_{i0})^2 + \sum_{angles} \frac{k_\theta}{2}(\theta_i - \theta_{i0})^2 + \sum_{torsions} \frac{V_n}{2}(1 + \cos(n\omega - \gamma))$$

$$+ \sum_{i=1}^{N} \sum_{j=i+1}^{N} \left(4\varepsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} + \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right] + \frac{q_i q_j}{4\pi\varepsilon_0 r_{ij}} \right)$$

The first two terms in this equation represent the chemical bonds and the atomic angles, which we approximate as springs (Hooke's law). The third term represents dihedral angles, which we represent using a sinusoidal function. The last term describes the nonbonded interactions. To calculate Van der Waals forces and electrostatic interactions, we use the Leonard-Jones potential and Coulomb's law, respectively.[27,28]

MD simulations are used extensively in drug discovery. For example, they can provide different protein conformations for use in ensemble docking.[24] After docking, MD simulations can determine binding-pose stability, as demonstrated by Kokubo's group,[29,30] which has used simulations to further discriminate ligand poses obtained by docking. MD simulations have also been used to find cryptic pockets, also known as hidden allosteric pockets. These pockets open the possibility of targeting proteins previously thought of as undruggable.[31,32] Although most cryptic pockets have been found by chance, and there is a growing effort focused on finding new proteins with cryptic pockets.[33]

Finally, another use for MD simulations in drug discovery is calculating binding free energies using methods like thermodynamic integration and free energy perturbation, among others. These computationally expensive methods are the most accurate in predicting how well a ligand binds a protein. The difference in free energy between two related molecules can be used to determine if a ligand modification will improve its binding to the target protein.[34] This method can be used to create relationships between the modification and binding affinity, generating a structure-activity relationship model.[35]

1.3 Conclusions

Here I briefly introduced the field of computer-aided drug design, focusing on methods that use the structure of the target protein (SBDD methods). CADD has accelerated drug discovery by providing protein models, minimizing the number of compounds that need to be experimentally tested, and generating structure-activity relationships.

Drug discovery may be in a paradigm change due to the advances in CADD methods, which are not only aiding in drug discovery but are driving this enterprise.[36,37] The goal of CADD will always be to deliver new and safer medicines fast and cost-effectively.

2.0 Sub-Pocket EXplorer (SubPEx): Leveraging weighted ensemble simulations to enhance the conformational sampling of binding-pocket conformations

2.1 Introduction

Drug discovery is a lengthy and expensive endeavor with research and development costs ranging from 161 million dollars to a staggering 4.54 billion dollars for new drugs.[1,2] Computational methods are used extensively in most steps of the drug discovery pipeline. Virtual screening can alleviate the high costs associated with early-stage hit identification. Standard virtual-screening methods dock flexible compounds into a single, rigid protein receptor. These methods are fast, but they do not consider all the conformational sub-states a protein can adopt. Ensemble docking seeks to overcome this limitation by docking candidate ligands into multiple protein conformations.[24,25,38,39] These protein conformations are often derived from brute-force molecular dynamics (MD) simulations. However, standard MD usually takes too long to thoroughly sample the entire conformational landscape, even with advances in computational resources like graphics processing units (GPUs) and parallel computing.[40,41] Energy landscapes typically have metastable states with large energy barriers separating them; crossing these barriers is a rare event that MD simulations have difficulty sampling. The scientific community has developed several methods to overcome this limitation, including methods to simplify the systems (coarse-graining), modify the energy landscape (accelerated MD, replica exchange MD, metadynamics), and focus computational efforts on rare-event sampling (transition path sampling, weighted ensemble). [42–45]

Huber and Kim[46] developed[†] the weighted ensemble (WE) path sampling method. This computational technique accelerates rare-event sampling by encouraging even sampling along a predetermined uni or multi-dimensional progress coordinate, which is usually divided into bins. The WE method involves two main steps; in the first step, several stochastic simulations (walkers), each carrying a statistical weight, are performed for a predefined interval (τ). In the second step, merging and splitting ensures even distribution in the progress coordinate space. The probabilities of the involved walkers are divided or added in the merging and splitting, respectively.

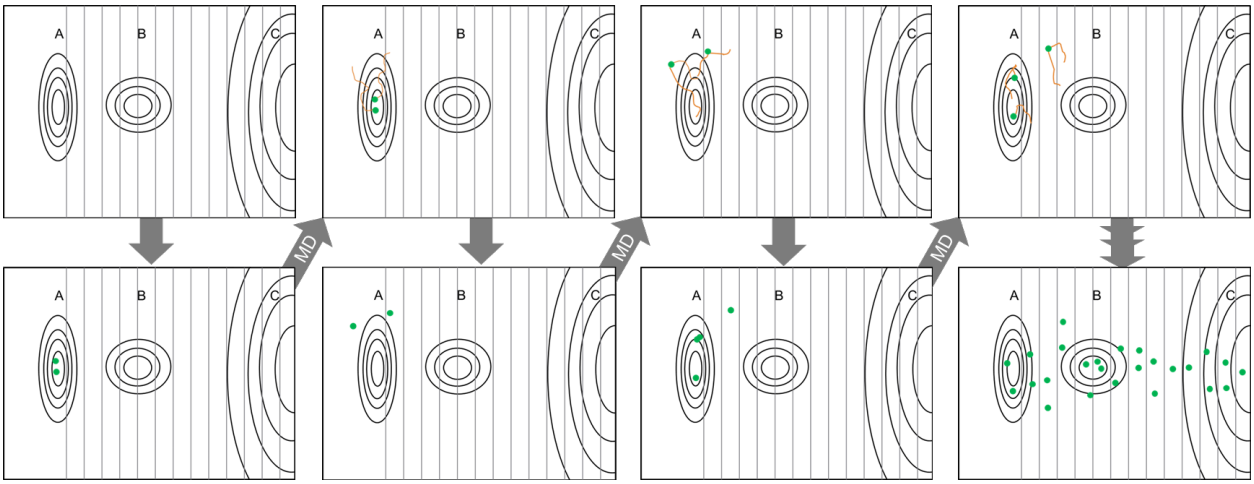


Figure 3. Graphical description of the weighted ensemble method. Depicted here is a weighted ensemble simulation that starts by populating the first bin in the progress-coordinate phase space with two walkers. After short MD simulations, one walker has crossed to another bin, so we repopulate each occupied bin with two walkers, each with half the probability of the parent simulation. Following another short MD simulation, we repopulate the third bin and merge walkers in the first bin. We continue to perform these two steps as needed.

[†] Some consider this a rediscovery of the splitting strategy developed in 1951 by Kahn and Harris.

In this work, I present the development of Sub-Pocket EXplorer (SubPEx), a novel tool to obtain ensembles of protein pocket conformations. To accelerate sampling, SubPEx uses WE as implemented in WESTPA, an open-source, highly scalable WE implementation.[47,48] Here, I describe the development of SubPEx, its progress coordinate, and how it compares to vanilla MD simulations. Also, I developed a clustering algorithm that clusters per generation. This clustering algorithm gives better and faster results than using every SubPEx simulation frame. Finally, I apply SubPEx to three proteins relevant to drug discovery: heat shock protein 90, influenza neuraminidase, and yeast hexokinase II.

2.2 Methods

2.2.1 Preparation of proteins for simulations

Proteins models were created from crystal structures downloaded from the Protein Data Bank.[49] For the ATP binding domain of heat shock protein 90 alpha (HSP90), I used PDB 5J2V[50]; for N1 Neuraminidase (NA), I used PDB 2HU4 and 2HTY[51] for closed and open conformations, respectively. I removed the bound ligand (oseltamivir) in the closed NA conformation. Finally, for *Saccharomyces cerevisiae* hexokinase II (ScHxk2p), I used PDB 1IG8.[52]

I added hydrogen atoms to each protein using the PDB2PQR[53] webserver (pH 7.0), which uses the PROPKA algorithm to optimize the hydrogen-bond network.[54] I used LEaP, part of the AmberTools18 package[55], to add a water box that extends in every direction for 10 Å beyond the protein. I also added Na+ or Cl- ions to neutralize the charge of the protein and added additional

ions to approximate a 150 mM NaCl solution. For the proteins, I used the Amber ff14SB force field[56], and for water, I used TIP3P.[57]

To obtain a protein/solute system without any steric clashes, I applied four rounds of minimization with 5000 steps each. I used either NAMD (versions 2.13 and 2.14)[58] or Amber from Amber[55] to perform this step. In the first minimization step, I allowed hydrogen atoms to be free. In the second, hydrogen atoms and water molecules. In the third, hydrogen atoms, water molecules, and protein side chains were allowed to relax. Finally, in the fourth step, all atoms were free. I followed this with an equilibration period of at least one ns.

For the simulations using NAMD, the equilibration was done serially, gradually relaxing the restraints on the backbone (1.00, 0.75, 0.50, 0.25, and 0.00 kcal/mol/Å^2, respectively). All the steps were done in the NPT ensemble; the first step used a one fs timestep, and the subsequent steps used a two fs timestep.

For the Amber simulations, I used a three-step equilibration. The first step was in the NVT ensemble for 10,000 steps with 1 kcal/mol/Å^2 constraints on backbone atoms. The second step is in the NPT ensemble, for one ns of total time using a two fs timestep and 1 kcal/mol/Å^2 constraints for backbone atoms. The last step is the same as the previous step but without constraints on the backbone.

2.2.2 Molecular dynamics and weighted ensemble simulations

To run the WE and vanilla molecular dynamics (MD) simulations, I used NAMD 2.13, NAMD 2.14, Amber18, and Amber20. I did not mix engines or versions; for the systems equilibrated with NAMD, I continued with NAMD, and for the systems equilibrated with Amber, I continued with Amber.

For each system, the vanilla MD simulations were run in the NPT ensemble, using a two fs timestep. For the SubPEx and the vanilla simulations, I used the Monte Carlo barostat with a pressure of 1.01325 bar (1 atm). Since WE simulations rely on stochastic processes, I implemented the Langevin thermostat with a collision frequency of 5.0 ps^{-1} and a target temperature of 310K for both types of simulations. I performed three production runs, one 500 ns and two 250 ns, for a total of 1 μs for each system. All systems were neutralized with Na^+ or Cl^-, and I added the same ions to approximate a 150 mM solution.

SubPEx uses WE to accelerate pocket conformational sampling. I used WESTPA 1.0 as the WE implementation.[47] WESTPA is compatible with any molecular dynamics engine, so I implemented two engines in SubPEx; I used the same engines as in the vanilla MD simulations, Amber and NAMD. For binning, I used the minimal adaptive binning algorithm developed by Torillo and coworkers.[59] All WE simulations were performed with a tau of 20 ps; the specifics of each simulation (*e.g.*, number of bins, walkers per bin) are specified for each WE simulation in the section where each simulation is discussed (Section 2.3.1, 2.3.3, and 2.3.4).

2.2.3 Analysis of simulations

Progress coordinate calculations were done using in-house scripts, I used MDAnalysis for all molecular data reading and manipulation. PCA analysis was done using the MDAnalysis python package version 1.0.0.[60,61] Clustering the points I used to fill the pocket was done using SciKit-Learn (version 0.22.1).[62]

Clustering of the MD simulation trajectories was performed with Amber20's CPPTRAJ[63] using hierarchical agglomerative clustering with average-linkage, which uses the average distance between members of two clusters to calculate the similarity between that pair of structures.[64]

2.3 Results and discussion

2.3.1 Development of SubPEx and its progress coordinate

We first hypothesized that we could use a pocket-similarity metric, the pocket-shape Jaccard Distance (JD), as the SubPEx progress coordinate. To calculate JD, the user provides the center of the pocket to be sampled, and the SubPEx algorithm creates a field of points (FOP) that fills the pocket. The procedure is similar to that used by the POVME algorithm.[65,66] It starts by creating a sphere of points centered at a user-defined point, this point should be at the center of the pocket the user wants to enhance sampling. To delete points that clash with any protein atom, I used a VdW radius of 2.6 Å. Next, it calculates and deletes all points outside the convex hull created by the Cα atoms of the protein. Finally, we cluster the FOP using DBScan to obtain the final FOP. This FOP is then compared to the FOP of a reference, and the JD is calculated as follows:

$$JD_i = 1 - \frac{\left|FOP_i \cap FOP_{ref}\right|}{\left|FOP_i \cup FOP_{ref}\right|}$$

The JD metric is degenerate, and we can break the degeneracy of JD by introducing a second progress coordinate, which is why I tested the SubPEx algorithm with a two-dimensional progress coordinate. The first dimension was JD, and the second was the pocket heavy-atoms RMSD (pRMSD). Heavy atoms in the pocket are obtained by an algorithm that takes every residue within the initial sphere of points. Then the user is encouraged to visually inspect and remove residues that are not part of the protein pocket.

I tested this 2D progress coordinate on the open 150-cavity conformation of influenza neuraminidase (see section 2.3.3), a viral protein that allows the influenza virion to leave its host

cell. The simulations had a τ of 20 ps, five independent trajectories ("walkers") per bin, 68 bins, and 75 total iterations. The maximal trajectory length, the total simulation time for a single walker from generation 1 to the last one, was 1.50 ns with 360.3 ns aggregate simulation time (concatenation of all the walkers in all the generations).

A comparison of the 2D NA SubPEx simulation and the three vanilla MD simulations (totaling 1 μs) can be seen in Figure 4 and Figure 5. The SubPEx simulation samples more of the progress-coordinate space than the vanilla simulations, especially in the lower pRMSD and JD region. However, the vanilla MD simulations sample more of the higher pRMSD and JD values. It is also noteworthy that the vanilla MD simulations sample around what seems to be an energy minimum (in this particular space). The SubPEx simulation is not stuck at the minimum.

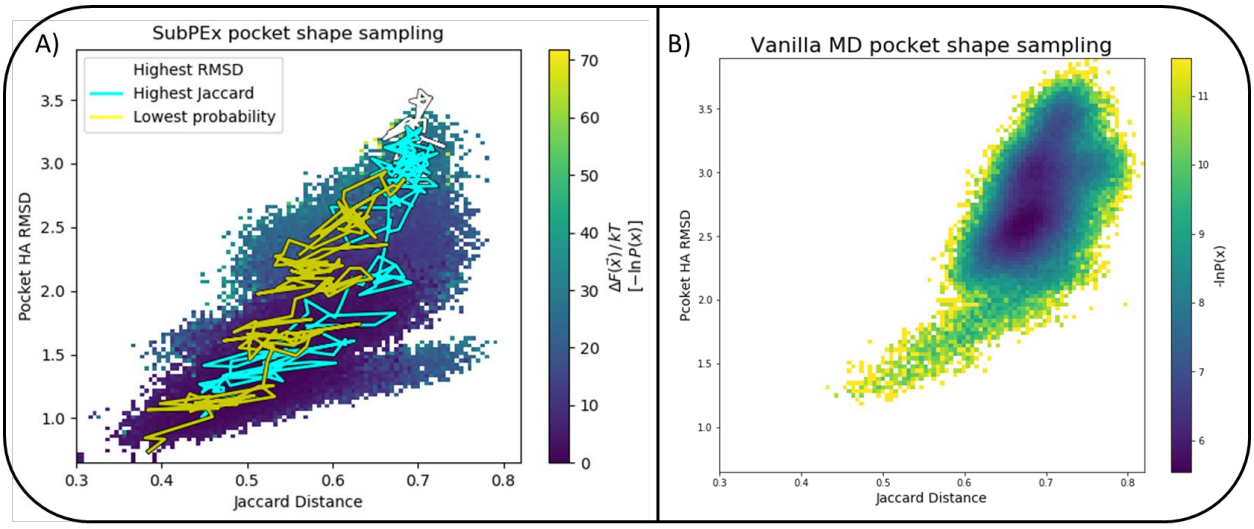


Figure 4. A) Two-dimensional probability distribution as a function of the JD (x-axis) and pRMSD (y-axis). In a yellow trace, I show the path that the lowest probability walker took; in blue and white, the walker with the highest JD and pRMSD, respectively. The cumulative simulation time is 360.3 ns. B) Two-dimensional probability distribution as a function of the JD (x-axis) and pRMSD (y-axis) of the vanilla MD simulations (total simulation time 1 μs), with the probability shown as counts (colors are inverted compared to A).

The violin plots shown in Figure 5 could explain why the SubPEx simulations do not adequately sample the higher pocket RMSD conformations in this case. The SubPEx simulations seem to be neglecting backbone sampling; by not allowing some backbone flexibility, the pocket atoms may be too constricted, limiting pocket sampling. An evident advantage of the SubPEx simulation is that the pocket conformational space that SubPEx samples, it does thoroughly. This thoroughness is observed by the pRMSD violin plot and the even distribution of pRMSD values observed for the SubPEx simulation.

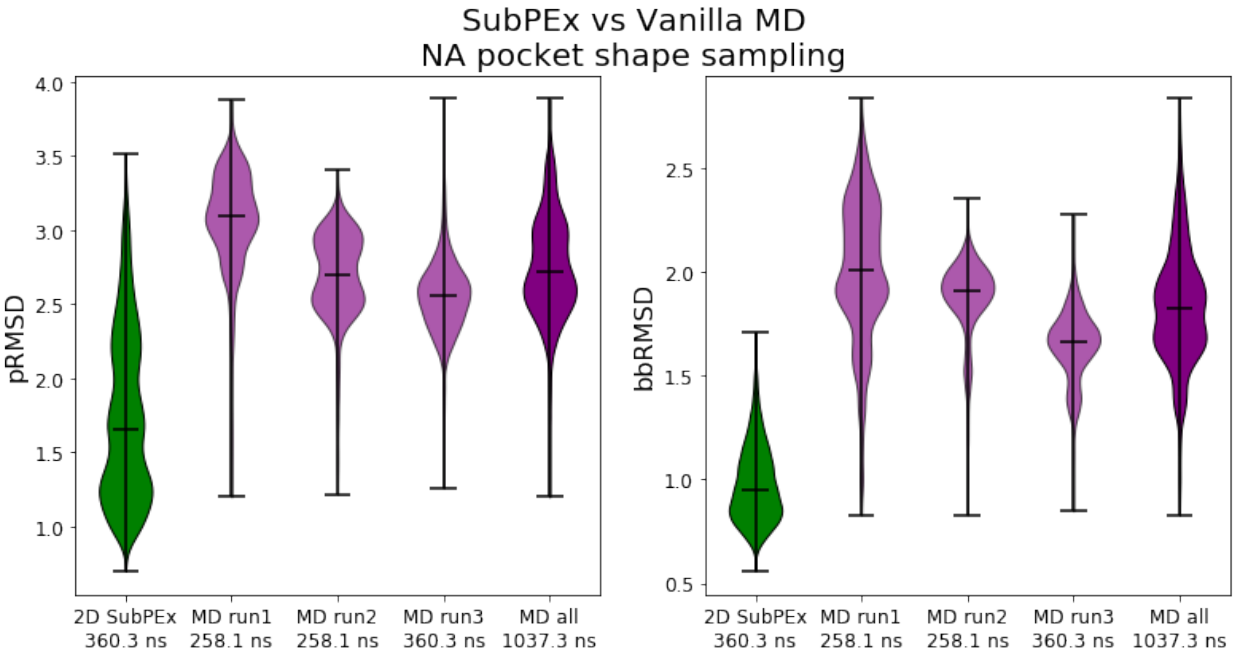


Figure 5. Violin plots of NA pocket heavy-atom RMSD (left) and backbone RMSD (right) for SubPEx and vanilla MD NA simulations. The vanilla MD simulations are presented in purple. In dark purple, I present the concatenation of the three MD simulations. Only the first 360.3 ns of the third vanilla MD simulation is included in this analysis. The other two MD simulations are presented without truncation since they are shorter than the 360.3 ns of the 2D-SubPEx simulation.

One of the most critical pieces of information the WESTPA algorithm provides, the probability of the walkers, is completely lost on all the violin plots presented in this chapter. WE is a demonstrated method that accelerates sampling of rare events and conformational sampling,[67–

[69] but to reach meaningful state probabilities these WE simulations have to equilibrate. All the simulations I present in this chapter have not reached equilibrium, which is why the probabilities each state has are not representative of reality. One thing I can do, is to reweight the probabilities, which is stated in the future directions section, and obtain probability-weighted violin plots. However, the project aims to study the extent of conformational sampling, not the kinetics or the thermodynamics of the process, which is why for the moment, we can safely ignore the probabilities.

I also applied SubPEx to the ATP binding domain of human heat shock protein 90 (HSP90), chosen because of its smaller size, the many structures deposited in the Protein Data Bank (260 structures with 100% sequence identity as of July 15th, 2022), and its relevance to cancer therapy.[70,71] HSP90 is a molecular chaperon that plays a central role in many cellular processes, including cell cycle control and survival. It is one of the most abundant proteins in the cytosol and is involved in maintaining cellular homeostasis. It is a frequently targeted protein in cancer drug discovery because its overexpression is an important factor in tumorigenesis.[70–73]

I began with an *apo* HSP90 (PDB 5J2V) structure. After minimization and equilibration using Amber20 (see methods section), I ran three vanilla MD simulations and a 2D-SubPEx simulation with the same parameters as the previous NA simulation, except I used three walkers per bin instead of the five. I ran this SubPEx simulation for 32 generations for a total cumulative simulation time of 104.52 ns and a maximal trajectory length of 0.64 ns.

When comparing SubPEx and vanilla simulations with the same cumulative simulation time, we see that SubPEx samples the pocket conformational space, per pRMSD, roughly the same as the three vanilla simulations. The violin plots in Figure 6 show that the SubPEx simulation samples more pocket diversity (per pRMSD) than the vanilla MD simulations but does not reach

high pRMSD values as does the second MD production run. As with the previously discussed NA simulation, SubPEx is still sampling low backbone RMSD (bbRMSD) values, limiting which pocket conformations are obtained in the simulation.

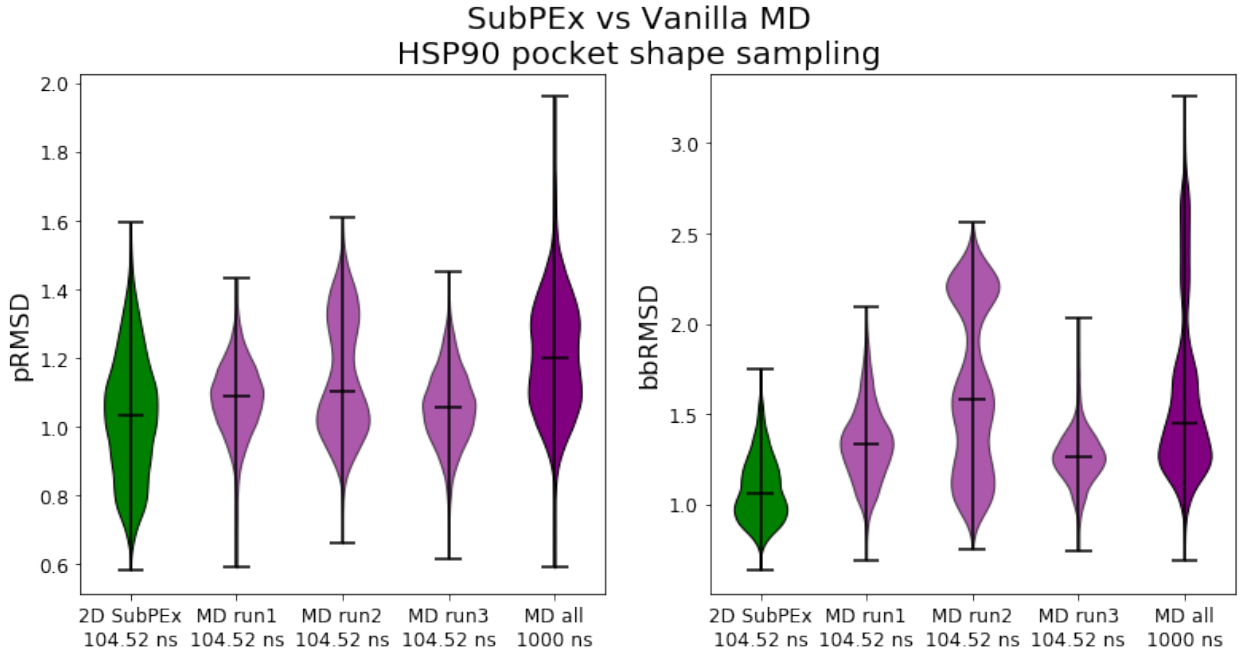


Figure 6. Violin plots of *HSP90* pocket RMSD and backbone RMSD for a 2D SubPEx simulation and three production runs of vanilla MD simulations. I present data of only the first 104.52 ns for each vanilla MD simulation.

To compensate for the hyper-focusing on the initial backbone conformation observed in both SubPEx simulations, I developed a new progress coordinate: the composite RMSD (cRMSD). My new hypothesis is that if we include some backbone flexibility into the progress coordinate, we can better sample the accessible pocket conformations while still focusing on pocket conformational sampling. I expect that some pocket conformations are going to be more accessible with modest backbone rearrangements. I defined the composite RMSD as:

$$cRMSD = pRMSD + \sigma \times bbRMSD$$

where bbRMSD is the backbone RMSD of the whole protein and σ is a proportionality constant (the percentage of backbone atoms not in the pocket divided by two). This constant was added to continue to focus SubPEx sampling on the pocket while introducing some of the whole-protein backbone dynamics.

To test the new progress coordinate, I set up three SubPEx simulations with one-dimensional progress coordinates (JD, pRMSD, and cRMSD) and three vanilla MD simulations (totaling 1 μs of simulation time). All the simulations use NAMD. I calculated pRMSD, cRMSD, and JD, for each SubPEx simulation. These metrics were calculated to be used as the progress coordinate or as auxiliary data, which can then be used to compare the simulations. Each SubPEx simulation had 19 bins using the MAB scheme,[59] had three walkers per bin, and was run for 50 generations using a τ of 20 ps. The runs had 46.98, 49.62, and 47.88 ns of cumulative simulation time and a maximal trajectory length of one ns. Figure 7 shows the probability distributions per generation (iteration) for the three simulations.

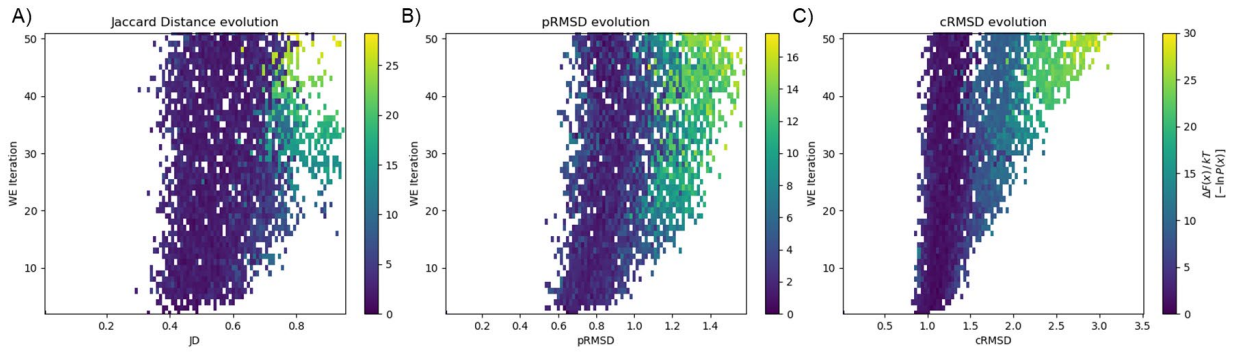


Figure 7. Probability distribution plots for the three 1D SubPEx HSP90 simulations. A) Simulation with JD as the progress coordinate. B) Simulation with pRMSD as the progress coordinate. C) Simulation with cRMSD as the progress coordinate. Note that the X axes and the color scheme do not share the same scale.

These plots show the sampling of the JD simulation is stalled; only a few walkers reach higher JD values, with most of the probability staying in the range of ~0.4-0.65. In contrast, in the pRMSD simulation, more walkers sample higher progress-coordinate values. Finally, in the

cRMSD progress coordinate, many walkers reach higher progress-coordinate values, and the probabilities are more spread out compared to the other simulations.

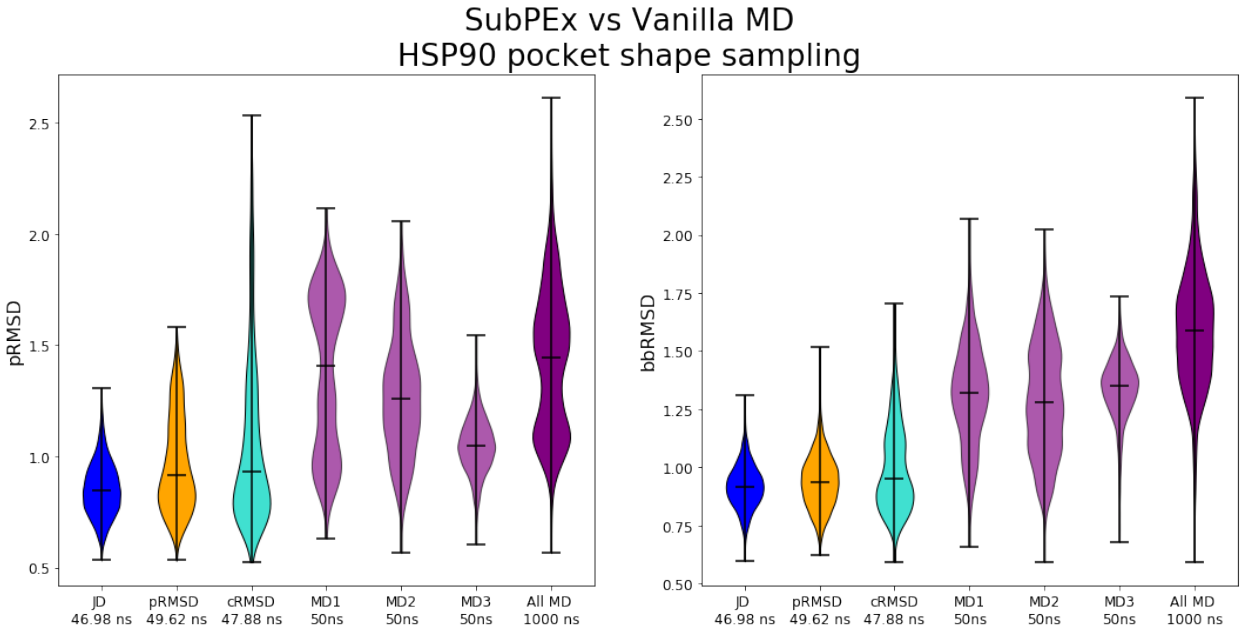


Figure 8. Violin plots comparing SubPEx HSP90 simulations with vanilla MD simulations. In blue, orange, and turquoise, we have the SubPEx simulations run using the JD, pRMSD, and cRMSD progress coordinates, respectively. In purple, three vanilla MD simulations up to 50 ns, with darker purple being the concatenation of all the frames of the vanilla simulations for a total of 1 μs of simulation time.

To compare pocket sampling between simulations, I created violin plots showing pRMSD and bbRMSD for the three SubPEx simulations and the three vanilla simulations. These plots show that the cRMSD SubPEx simulation outperforms all other simulations according to the pRMSD metric, both in maximum pRMSD value and sampling distribution. It almost reaches higher pRMSD values than even the 1 μs vanilla MD simulation. In comparison, the cRMSD samples more protein backbone conformations but still focuses conformational sampling on the pocket, as expected. As in the previous simulations, the JD and the pRMSD simulations were hyper-focused on the pocket with no enhanced backbone sampling. Given how well the cRMSD SubPEx

simulation performed, I extended it for 50 more generations (accumulative simulation time of 103.02 ns, shown in Figure 9 and Figure 10).

Finally, I explored whether a two-dimensional progress coordinate consisting of bbRMSD and pRMSD could outperform the new composite progress coordinate. This 2D SubPEx simulation also used the MAB binning scheme (108 bins). I ran the 2D SubPEx simulation with three walkers per bin, a τ of 20 ps, and 73 generations, for an accumulative simulation time of 253.74 ns and a maximal trajectory length of 1.46 ns.

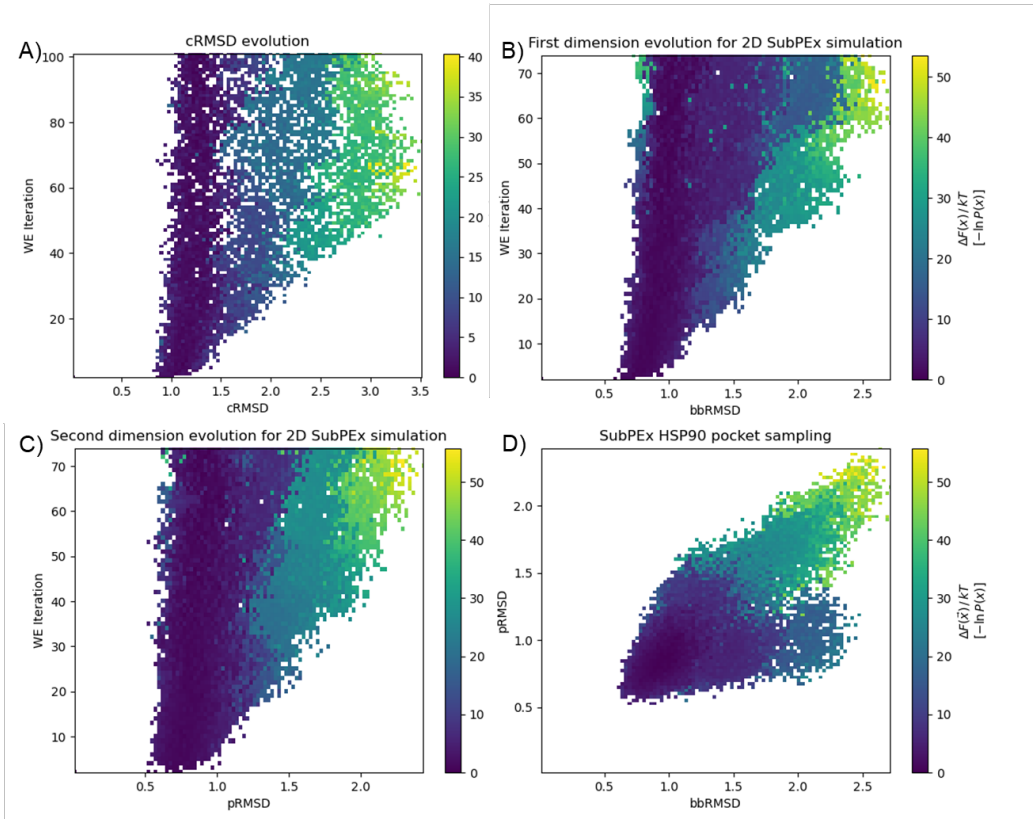


Figure 9. Probability distribution plots of SubPEx HSP90 simulations. A) Probability distribution per generation, cRMSD simulation (accumulated simulation time 102.6 ns). B) Probability distributions per generation for the first dimension (bbRMSD) of the 2D SubPEx simulation. C) Probability distributions per generation for the second dimension (pRMSD) of the 2D SubPEx simulation. D) Two-dimensional probability distributions as a function of bbRMSD and pRMSD for the 2D SubPEx simulation (accumulated simulation time 102.6 ns).

The probability distribution plots comparing the 1D cRMSD and the 2D bbRMSD/pRMSD SubPEx simulations (Figure 9) show that the 1D SubPEx simulation outperforms the 2D simulation, even when the simulation is less than half the cumulative simulation time (103.02 vs. 253.74 ns). The difference is even more striking when we compare simulations with roughly the same cumulative simulation time, as seen in Figure 10. In this plot, I show data from the first 31 generations of the 2D SubPEx simulation; separately, I present data from the whole simulation. I did this truncation of data because the 31-generation mark had about the same cumulative simulation time as the cRMSD SubPEx simulation. The cRMSD SubPEx simulation sampled higher pRMSD values compared to any other simulation. It also had a more even sampling distribution (*i.e.*, it sampled low and high pRMSD values).

Interestingly the 2D SubPEx simulation focused more on exploring backbone conformations than pocket conformations, as exemplified by their high bbRMSD values compared to all other simulations. In contrast, the simulation with the 2D progress coordinate falls short, both in magnitude and distribution of pRMSD values, compared to the vanilla MD simulations and the cRMSD SubPEx simulation. In comparison, the cRMSD simulation is still focused on sampling the pocket but allows for improved backbone conformational sampling.

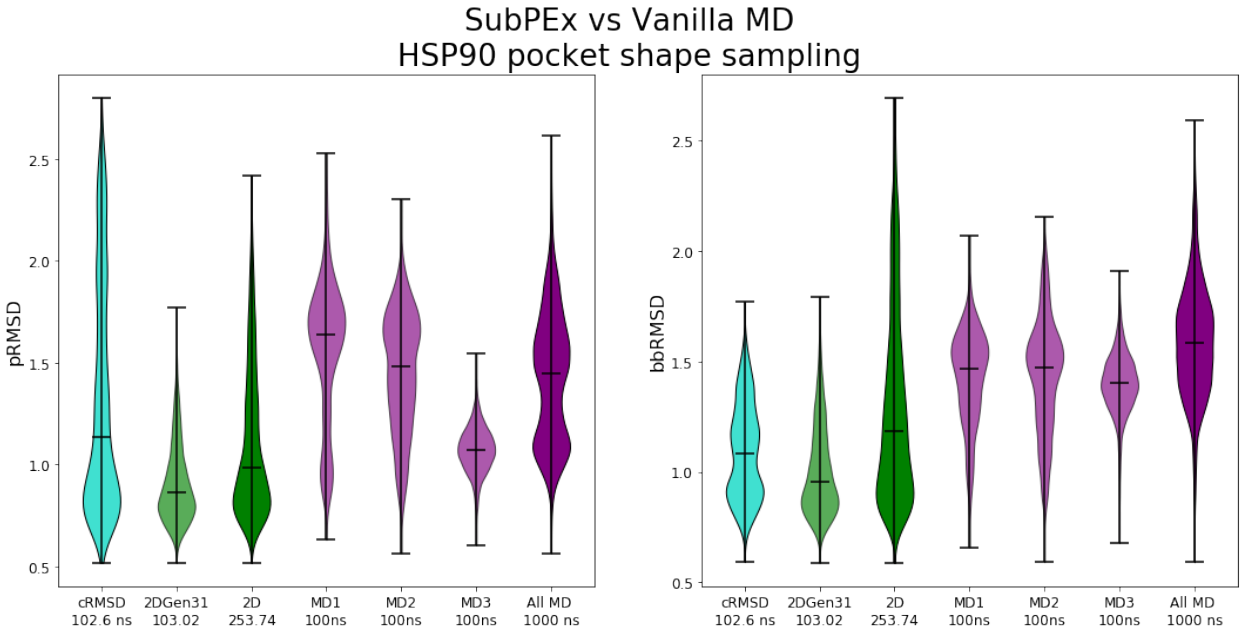


Figure 10. Violin plots comparing SubPEx HSP90 simulations with vanilla MD simulations. In turquoise, a SubPEx simulation using cRMSD as the progress coordinate. In green, the 2D SubPEx simulation with bbRMSD and pRMSD as the progress coordinates. Lighter green show results up to generation 31 (approximately the same simulation time as the cRMSD simulation). In purple, three vanilla MD simulations up to 100 ns, with darker purple being the concatenation of all the frames of the vanilla simulations.

Having demonstrated the pocket conformational sampling power of the cRMSD progress coordinate, I wanted to verify that we are sampling conformations that are attainable. *In lieu* of reweighing or running the simulations to equilibrium, I explored whether the HSP90 SubPEx simulations sampled conformations are similar to protein conformations in the PDB Data Bank. I performed principal component analysis (PCA) on pocket heavy atoms using the cRMSD SubPEx simulation, three truncated vanilla simulations, and a collection of 75 HSP90 crystal structures extracted from the PDB Data Bank. To calculate the PCs, I truncated the vanilla simulations to be the same length as the SubPEx simulation.

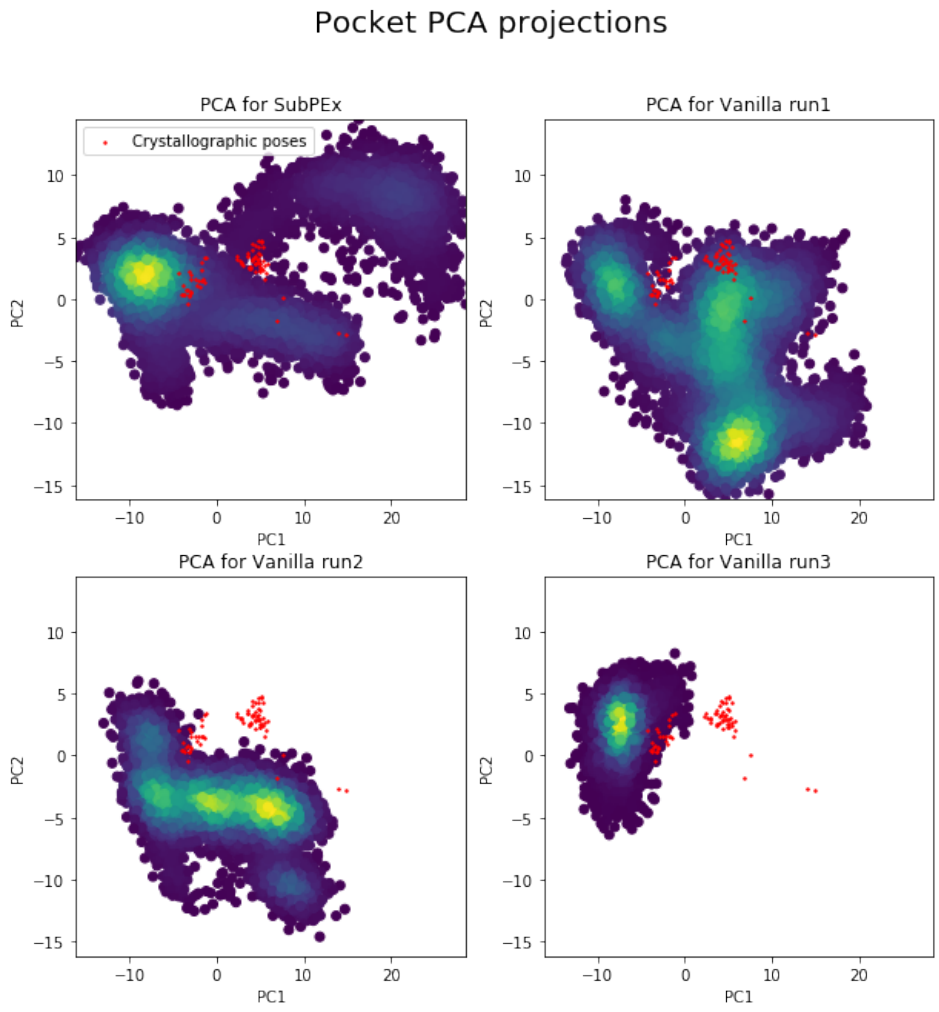


Figure 11. PC2 vs. PC1 plots of the HSP90 cRMSD SubPEx and three vanilla MD simulations. To ensure I used the same PC space between simulations, I performed a single PC analysis on a concatenated simulation with the SubPEx, the vanilla, and crystallographic structures.

Figure 11 shows the first and second PCs for all three vanilla MD simulations, the cRMSD SubPEx simulation, and the crystallographic structures (marked as red dots).

To confirm the results observed in the violin plots, I used the same PC analysis to calculate the percentage coverage of the PC space by each simulation. SubPEx samples the PC space more thoroughly than the vanilla MD simulations with ~47% coverage compared to the ~43, 25, or 11% coverage of the vanilla simulations.

To confirm which simulation shares more of the PC space with the crystallographic structures. I performed a PCA on the pocket atoms using only the SubPEx simulation and the first vanilla MD run. This analysis shows (Figure 12) that the SubPEx shares more PC space with the 75 crystallographic structures compared to the first vanilla MD run. These analyses taken together indicate that the SubPEx simulations sampled more conformations that are experimentally relevant compared to the vanilla MD simulations.

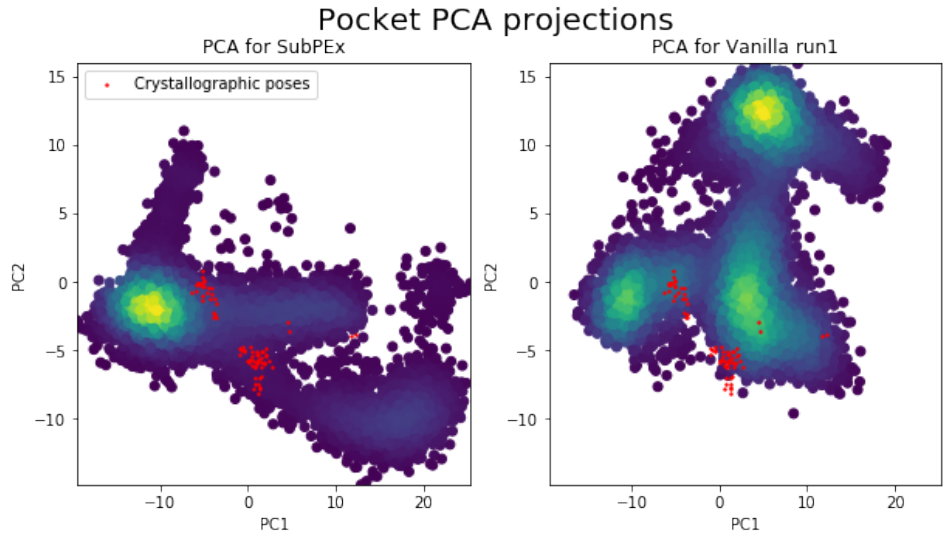


Figure 12 PC2 vs. PC1 plots of the cRMSD SubPEx and the first vanilla MD simulations. To ensure I used the same PC space between simulations, I performed a single PC analysis on a concatenated simulation with the SubPEx, the first vanilla MD, and the crystallographic structures.

2.3.2 Clustering of simulations

Having established that the SubPEx-sampled conformations are experimentally relevant according to PCA, I next explored how to extract meaningful conformations for later use in ensemble docking. Clustering is the logical approach for obtaining individual representative protein conformations. However, the confirmations obtained from SubPEx are not linearly

correlated, as are the conformations obtained by vanilla MD, and we should aim to use all of SubPEx's data. Using every frame in the SubPEx simulation frame for clustering can be troublesome because calculating the pair-wise distance matrix needed is computationally expensive in terms of memory and time.

To minimize the time it takes to calculate the matrix while still using every frame, I developed a script that clusters the data on each generation and then performs a final clustering with the centroids of the previously clustered data. This script considers the number of walkers in each generation, so I don't introduce bias in the final clustering results to any generation. The clustering focuses on the pocket conformations, and the algorithm I used is hierarchical agglomerative clustering using average linkage, as implemented in CPPTRAJ. The distance matrix calculation has a quadratic time complexity, meaning that calculating the distance matrices for the per generation algorithm will take substantially less cumulative time than the single distance matrix calculation using all the frames. Another advantage of the per generation algorithm is that each cluster's centroids are further apart when clustering per generation, as observed in the "all vs. all" pocket RMSD plots shown in Figure 13B. Both sides of this plot represent data from the SubPEx simulation; on the left, I present the clustering using the per generation algorithm, and on the right, I show clustering using all the frames of the SubPEx simulation.

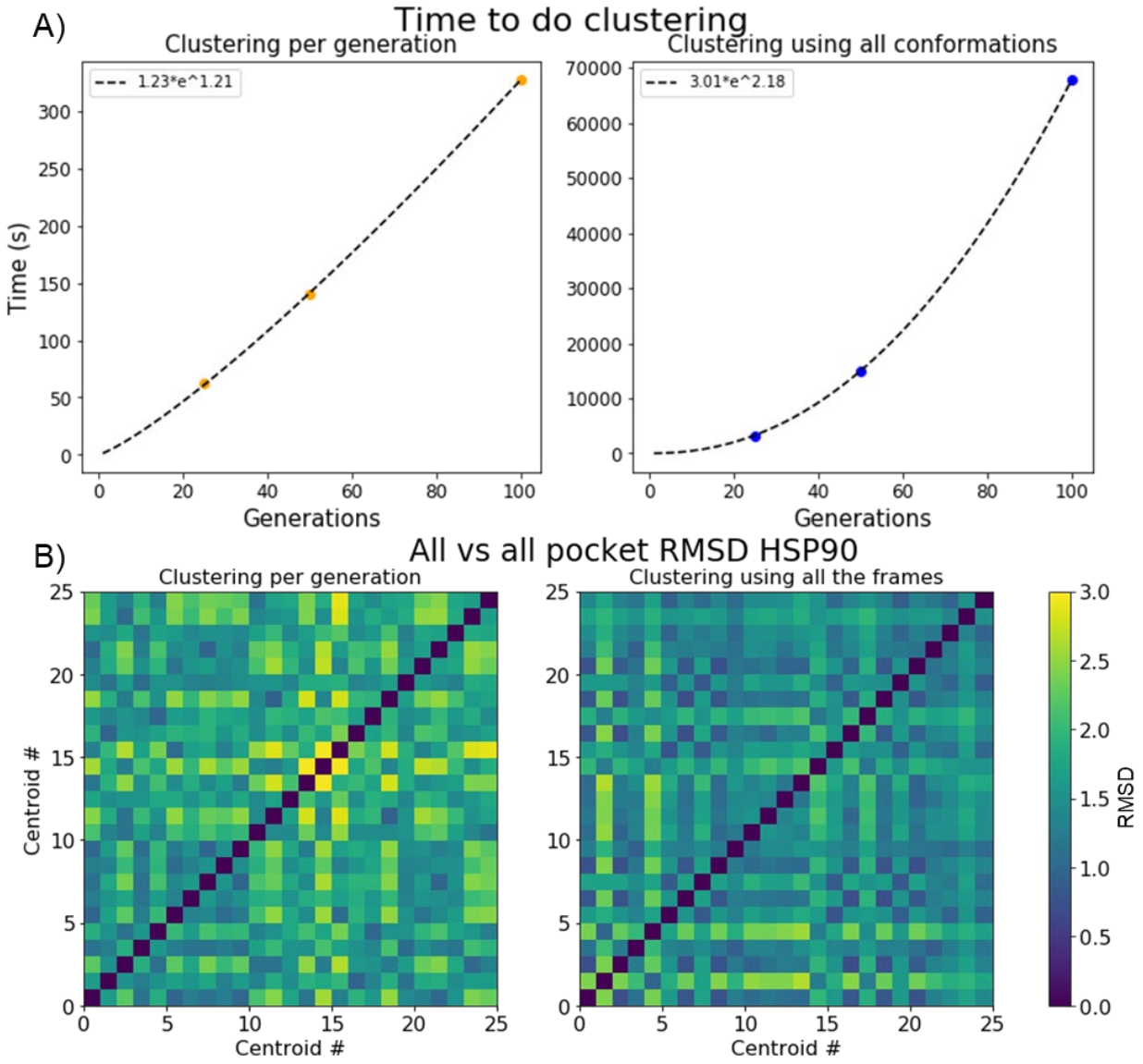


Figure 13. A) Plot of the time it takes to cluster using per generation clustering compared to clustering using all the frames. B) All vs. all pocket RMSD of the centroids obtained from clustering. Left) clustering per generation, right) clustering using every frame.

To further compare the SubPEx and vanilla MD simulations, I clustered the cRMSD HSP90 simulation using the per generation script and the vanilla simulations using every frame. Plots showing all vs. all backbone and pocket RMSD for the vanilla and SubPEx simulations are given in Figure 14. Comparing results from SubPEx and the first 102.6 ns of the third vanilla MD

production run (same cumulative time) shows that there is more backbone RMSD sampling in the SubPEx simulation (Figure 14). This better backbone sampling disappears when we compare SubPEx's backbone sampling to the full microsecond vanilla simulations. In these simulations, the MD cluster centroids sample more backbone conformational space. Comparing the pocket RMSD between simulations suggests that the SubPEx sampling is better. In the full μs MD simulation, we observe regions of low pRMSD, especially in the pRMSD of the first five clusters. Comparatively, although the SubPEx simulation has some similar clusters, it still has more diverse pocket shapes overall per the pRMSD metric.

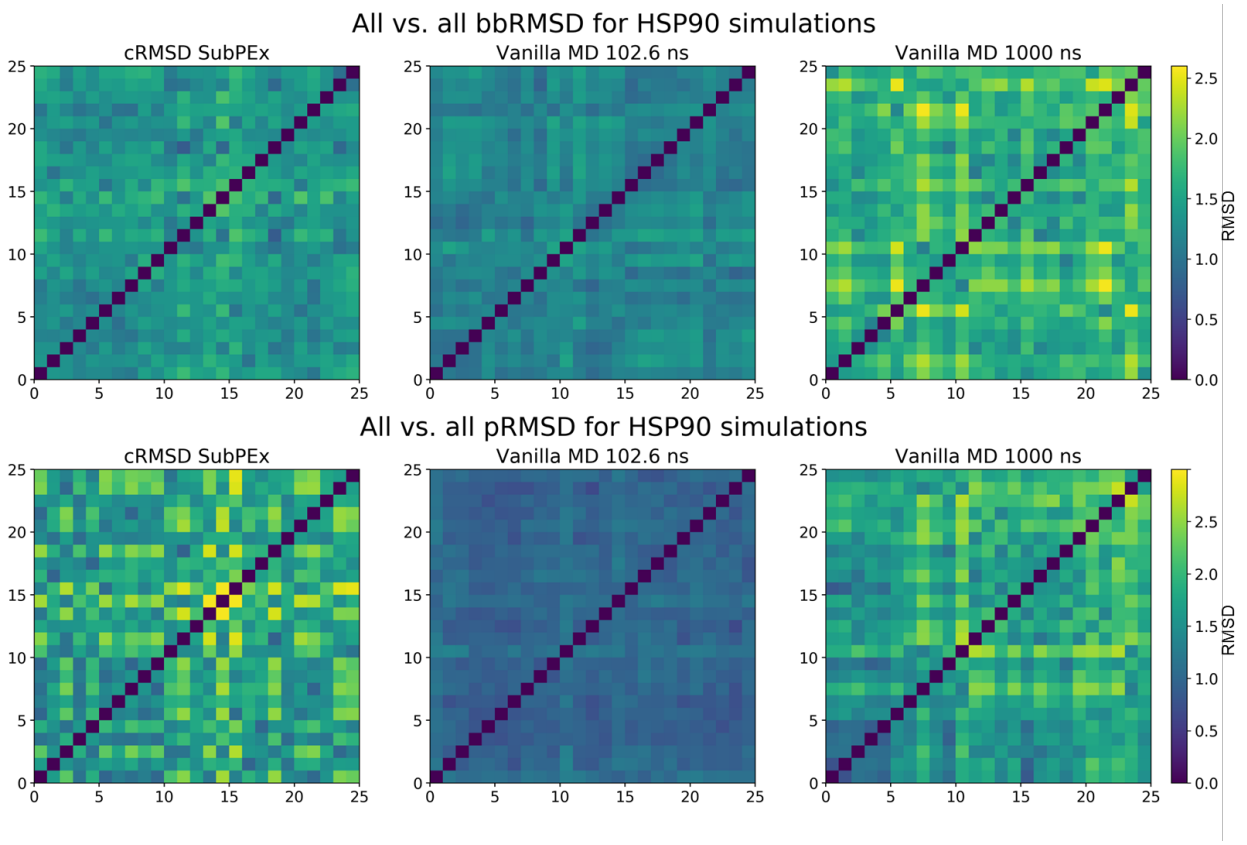


Figure 14. All vs. all RMSD plots for the cRMSD and vanilla MD HSP90 simulations. On top, I show backbone RMSD. On the bottom, I show pocket RMSD.

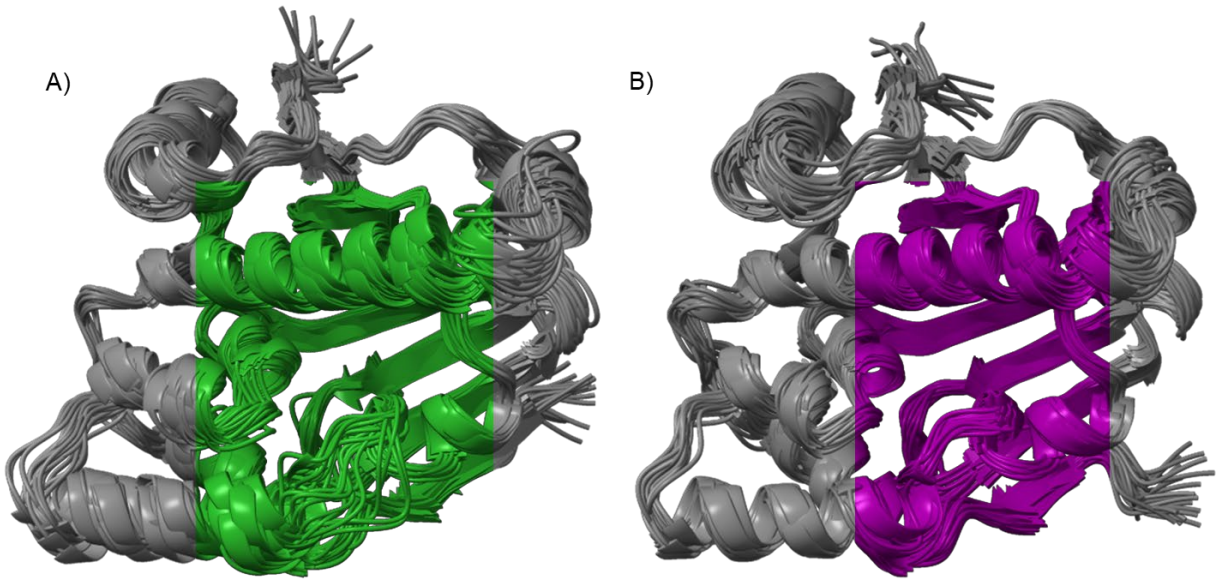


Figure 15. Superposition of HSP90 structures obtained from clustering SubPEx and vanilla MD simulations (same cumulative simulation time of 102.6 ns). A) Structures obtained by the cRMSD SubPEx simulation. B) Structures obtained by vanilla MD simulations.

These results show that the per-generation clustering algorithm is a fast algorithm to obtain distinct pocket conformations from WE simulations. I also demonstrate the conformational diversity observed in the HSP90 cRMSD SubPEx simulations seen in the all vs. all pRMSD plots of the centroids of the clusters obtained by the clustering algorithm.

2.3.3 Neuraminidase pocket sampling

To further test the SubPEx algorithm, I performed SubPEx and vanilla MD simulations on neuraminidase, a protein important in influenza's infection cycle. Influenza is a seasonal viral pathogen that, even though most people recover after a couple of days, influenza still kills between 290,000 to 650,000 people worldwide.[74] Viral replication ends with the budding and release of the viral entity. This release is mediated by neuraminidase (NA), the enzyme responsible for breaking

the sialic-acid linkages between viral hemagglutinin and the infected cell's membrane surface.[75–77] NA proteins have nine serotypes (N1 to N9), which can be divided into two groups according to their sequence. The main structural difference between both groups is the presence or absence of an extra cavity in the active site.[78,79] This cavity, which has been actively exploited for drug development,[80] is formed when the flexibility of a loop, the 150-loop, is increased by breaking the D147-H/R150 salt bridge.[81]

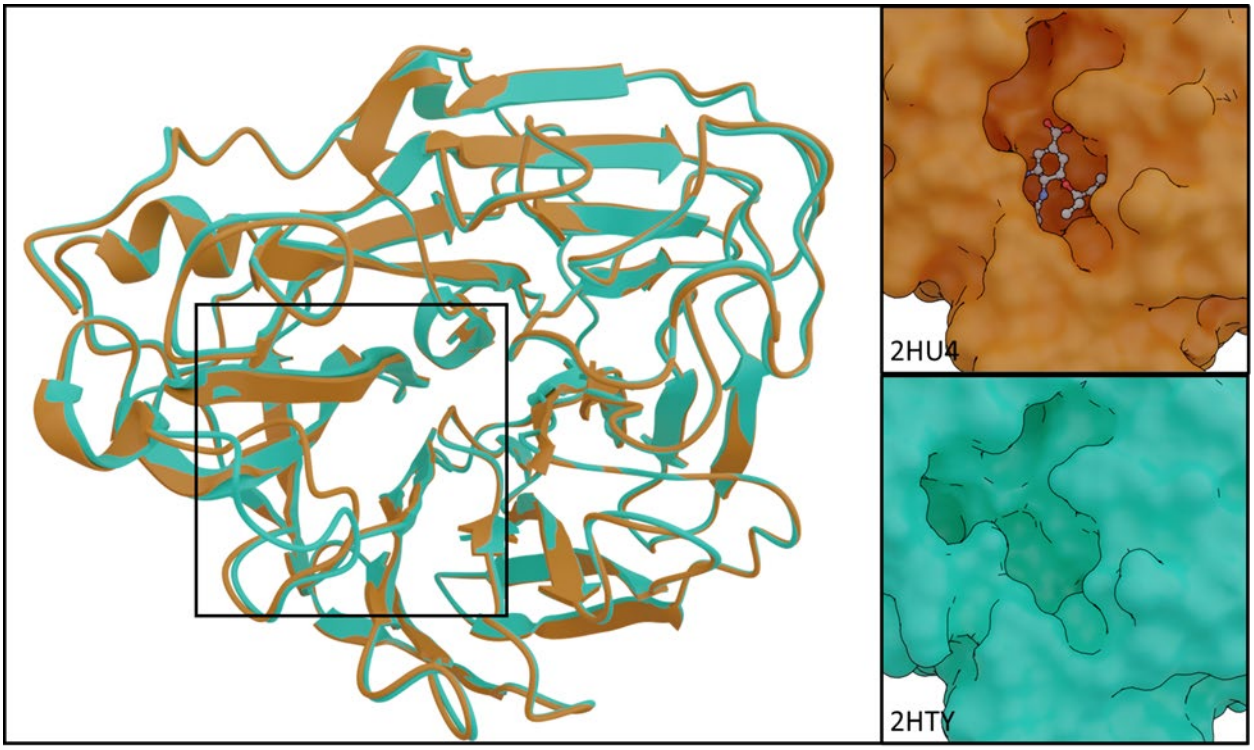


Figure 16. Superposed structures of neuraminidase with the 150-cavity in its closed (PDB 2HU4, orange) and open (PDB 2HTY, cyan) conformations.

To test SubPEx on this known flexible pocket, I prepared two NA systems for SubPEx and vanilla MD simulations. One starts from the open 150-cavity NA structure (PDB 2HTY:A, shown in cyan), and the second starts from the closed 150-cavity NA structure (PDB 2HU4:A, shown in orange). I used Amber to run the simulations for both systems (see Methods section).

Comparing the 1D cRMSD SubPEx and vanilla MD simulations starting from the open NA conformation shows that the SubPEx simulation samples higher pRMSD (Figure 17). Especially notable is the low bbRMSD sampled for the protein, demonstrating that we are focusing on the pocket diversity, not whole protein dynamics.

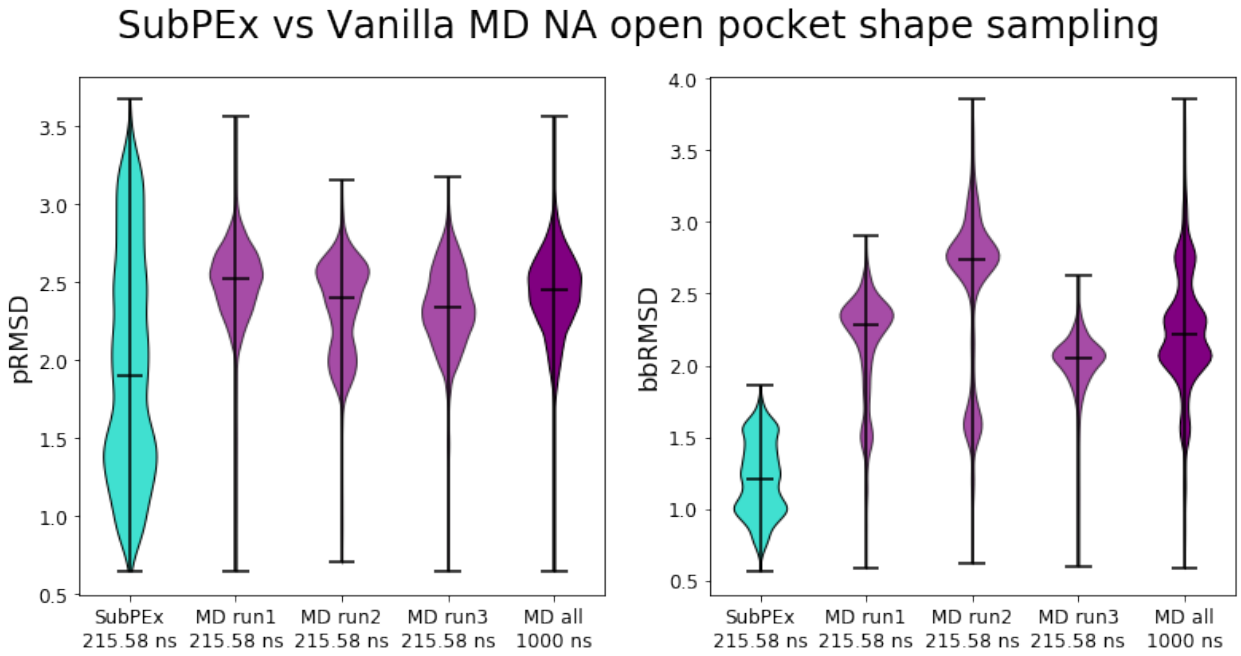


Figure 17. Violin plots comparing SubPEx NA simulations with vanilla MD simulations. In turquoise, SubPEx simulation using cRMSD as the progress coordinate, starting from the open conformation. In purple, three vanilla MD simulations using the first 215.58 ns, with darker purple being the concatenation of all the vanilla simulation frames.

When running the simulations starting from the closed conformation, the SubPEx simulations did not reach the same high pRMSD values as the vanilla simulations. I analyzed the vanilla trajectories and noticed that the system's equilibration was only achieved after ~50 ns for two of the three simulations. This result explains why SubPEx was not able to keep up with the vanilla simulations. SubPEx is spending its sampling power to equilibrate the closed system.

A simple analogy can explain this phenomenon; let us imagine that the vanilla MD is a single skier going down a hill, while SubPEx is like multiple skiers tied together trying to reach

the bottom of the mountain. The single skier will not be constrained and will arrive at the bottom of the hill (energy minima) more directly than the multiple skiers. Of course, the tied skiers will cover more terrain while going downhill compared to the single skier. This better sampling for low pRMSD values is observed in the distribution of the SubPEx simulation of NA starting from the closed conformation (data not shown).

2.3.4 Hexokinase II (Hxk2) results

To test SubPEx on a system that undergoes large domain rearrangements when binding a ligand, I ran a SubPEx simulation on *Saccharomyces cerevisiae* hexokinase II, a protein involved in glucose catabolism. In mammals, there are four hexokinases, all with a role in glucose metabolism. These enzymes begin the glucose catabolism when they phosphorylate glucose at its sixth position.[82,83] Cancerous cells have high energy requirements due to their unregulated cell growth and proliferation. These cells fulfill their energy requirements by increased glycolysis, known as the Warburg effect.[84–86] The increase in glycolysis can be achieved by overexpression of hexokinase 2; this overexpression is usually observed in cancerous cells. This makes Hxk2 an attractive target for novel anticancer therapeutics.

Using the *Saccharomyces cerevisiae* Hxk2 (*Sc*Hxk2p PDB 1IG8[52]) to model hexokinase behavior, I ran a SubPEx simulation and three vanilla MD simulations. Structurally, *Sc*Hxk2 contains two domains, a large and a small domain. These domains must come together to bind glucose. Results comparing the SubPEx and vanilla simulations are presented in Figure 18.

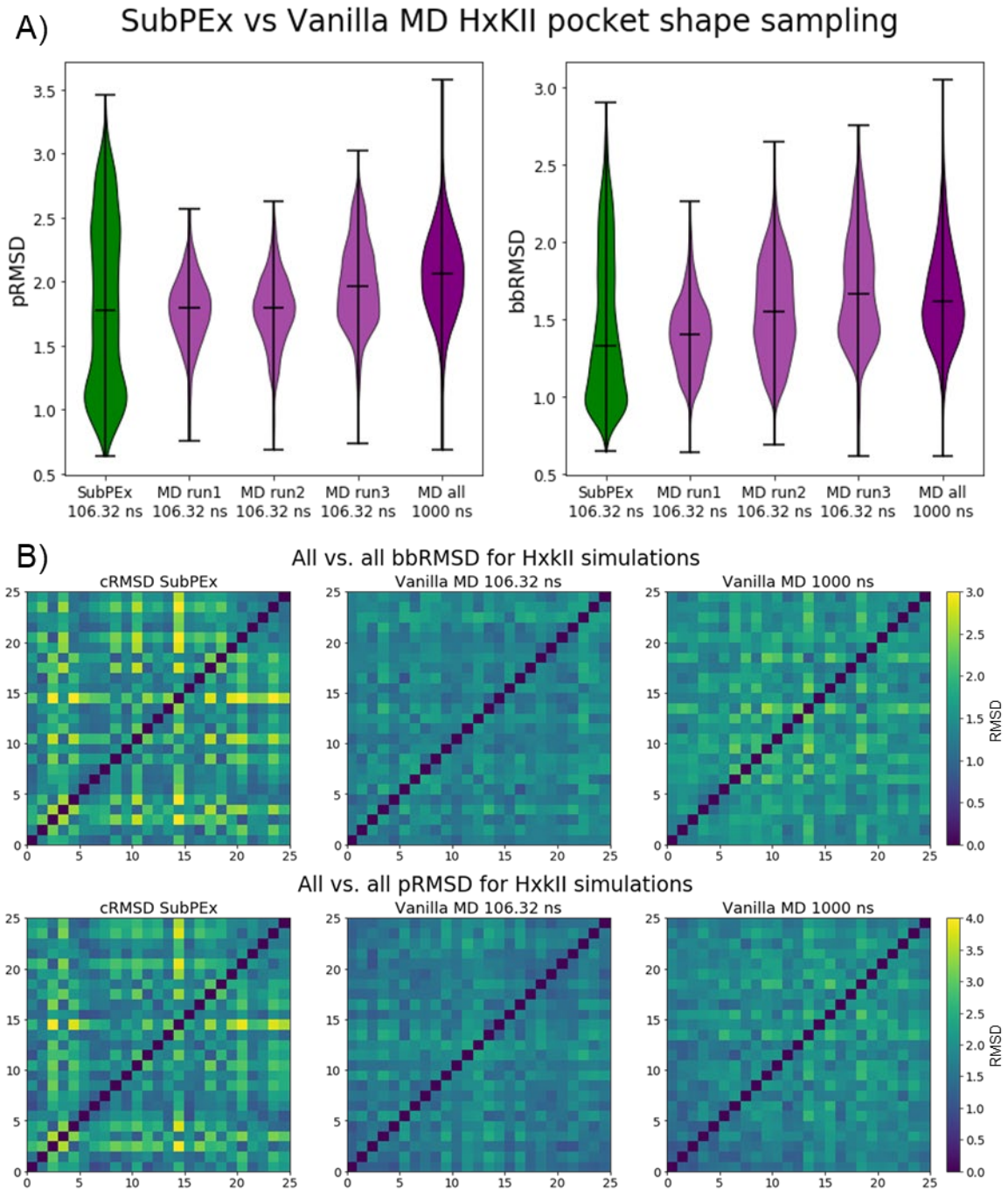


Figure 18. Hxk2 protein pocket sampling by SubPEx and vanilla MD. A) Violin plots depicting pocket and backbone sampling. B) All vs. all RMSD of clustered protein conformations for vanilla MD and SubPEx.

Especially telling of SubPEx's sampling capabilities is the comparison of both ensembles of protein conformations to crystallographic structures depicting hexokinase in its closed

conformation and (*Kluyveromyces lactis* hexokinase I, PDB 3O8M[87]) and open conformations (PDB 1IG8). The SubPEx protein ensemble is more similar to the closed conformation compared to the vanilla simulations (Figure 19), exemplifying how even when substantial conformational changes need to happen, SubPEx performs well.

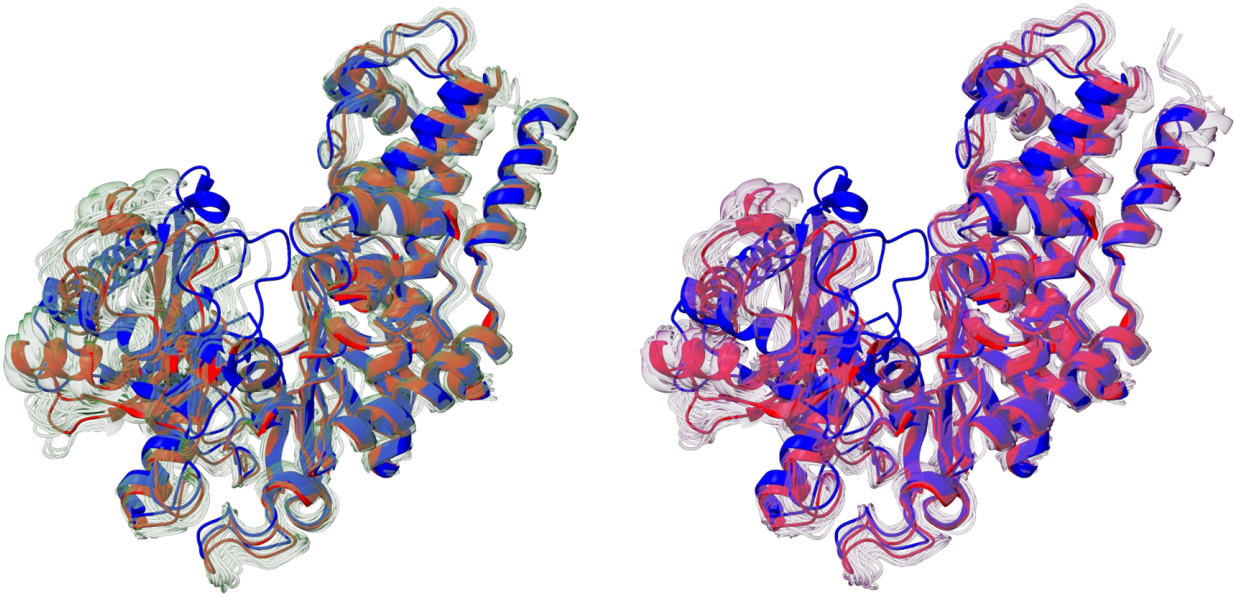


Figure 19. *Sc*Hxk2p conformational diversity as obtained by SubPEx and vanilla MD. In red, I show the starting conformation for the simulations (PDB 1IG8), and in blue, the closed hexokinase conformation (PDB 3O8M).

2.4 Conclusions

In this chapter, I discussed the development of SubPEx, a tool that improves pocket conformational sampling. I showed how incorporating some of the backbone's flexibility into the progress coordinate increases pocket conformational sampling without sacrificing SubPEx's computational focus on the pocket. With PC analysis, I showed that the SubPEx sampling includes

conformations observed experimentally. I also developed a clustering algorithm for SubPEx simulations that helps users efficiently obtain diverse pocket conformations for their later use in ensemble docking.

I demonstrated SubPEx efficacy using three systems. HSP90 simulations showed that the cRMSD progress coordinate allows for fast pocket sampling. Neuraminidase has a flexible loop and was a good candidate for SubPEx. NA starting from the open conformation showed better sampling compared to vanilla MD. Meanwhile, the NA system starting from the closed conformation demonstrates the importance of having a totally equilibrated system. If the system is not equilibrated, SubPEx will instead thoroughly sample the path to the energy minima, limiting the pocket sampling. To ensure an equilibrated state has been reached, I recommend autocorrelation analysis[88] or clustering analysis to monitor the number of clusters.[89] The last system I tested is *Sc*Hxk2p; this protein has a significant domain rearrangement when binding glucose. SubPEx captured this rearrangement, including both the closed and open conformations.

In conclusion, I show that SubPEx can accelerate the sampling of different protein pocket conformations compared to vanilla MD. This tool will help computational and medicinal chemists better incorporate protein flexibility into the drug discovery process. Specifically, with SubPEx one can obtain an ensemble of protein conformations with diverse pocket shapes in a time efficient. These protein conformations can then be used in ensemble docking.

2.5 Acknowledgments

I thank the University of Pittsburgh's Center for Research Computing and XSEDE (bio200078) for computational resources. I also thank the National Institute of Health

(1R01GM132353-01A1) for the support. I thank Kim F. Wong for help with debugging. I also thank Kevin Cassidy, Roshni Bhatt, and the Weighted Ensemble community for helpful discussions.

3.0 Novel mutation in hexokinase 2 confers resistance to 2-deoxyglucose by altering protein dynamics

This work has been published and can be found at:

Hellemann E*, Walker JL*, Lesko MA*, Chandrashekarappa DG, Schmidt MC, O'Donnell AF, Durrant JD. (2022) Novel mutation in hexokinase 2 confers resistance to 2-deoxyglucose by altering protein dynamics. PLOS Computational Biology 18(3): e1009929. https://doi.org/10.1371/journal.pcbi.1009929.*equal contribution.[90]

This work is a combination of experimental and computational methods. I did most of the published computational work, and I repeated or did not include in this chapter the analyses I did not personally perform. The physical experiments were performed mostly by Mitchell Lesko and Jennifer Walker. I have added a summary of the experimental results to give context to the computational results.

3.1 Introduction

Glucose is a molecule central to sustaining life and is used as an energy source and a building block for biosynthesis. After glucose intake by the cell, the sugar is converted into glucose-6-phosphate (Glc-6P), which is further transformed by anaerobic fermentation, aerobic oxidative phosphorylation, or the pentose-phosphate pathway. This first glucose transformation is catalyzed by a group of enzymes called hexokinases. Hexokinase 2 is the most active of the hexokinases[82] and is often upregulated in cancerous cells.[91,92]

Cancerous cells require more energy and metabolites due to their unregulated cell growth and proliferation. Tumors fulfill their high energy requirements by metabolic reprograming; specifically, cancerous cells rely more on glycolysis than healthy cells.[84,85] This effect was first described by Otto Warburg and is known as the Warburg effect.[93] It is not entirely understood what the benefits are of the reliance in glycolysis, given that cancerous cells do not fully oxidize the products of glycolysis. One theory is that oxygen supply is limited at the core of a growing tumor under rapid ATP generation, forcing the cell to rely on fermentation for ATP synthesis, rather than oxidation.[84,91]

Poor prognosis in several cancer types is associated with the upregulation of hexokinase II. In humans, Hexokinase 2 (*hs*Hk2) can bind mitochondria through its N-terminal helical domain. This association allows *hs*Hk2 to interact with the voltage-dependent anion channel (VDAC), placing *hs*Hk2 in a prime position to affect the apoptosis cycle. VDAC is responsible to release pro-apoptotic proteins, that are stored in the mitochondria, into the cytoplasm, but when VDAC interacts with *hs*Hk2 it interferes with the release.[94]

Structurally, *Saccharomyces cerevisiae* hexokinase II (*Sc*Hxk2p, from now on Hxk2) adopts a palm-shaped α/β fold with two subdomains: a large and a small subdomain. The protein starts in the so-called open conformation, with its enzymatic cleft accessible to substrates and water. When glucose binds, it induces a large conformational change where the two subdomains rotate relative to each other. This movement envelops or embraces glucose, making the inter-domain crevice inaccessible to additional substrates.[95–98] After glucose binding, ATP is able to bind Hxk2, priming the 6-hydroxy-methyl group to be acted upon by aspartate 211, the catalytic residue, and abstract the hydrogen atom in the hydroxy group.

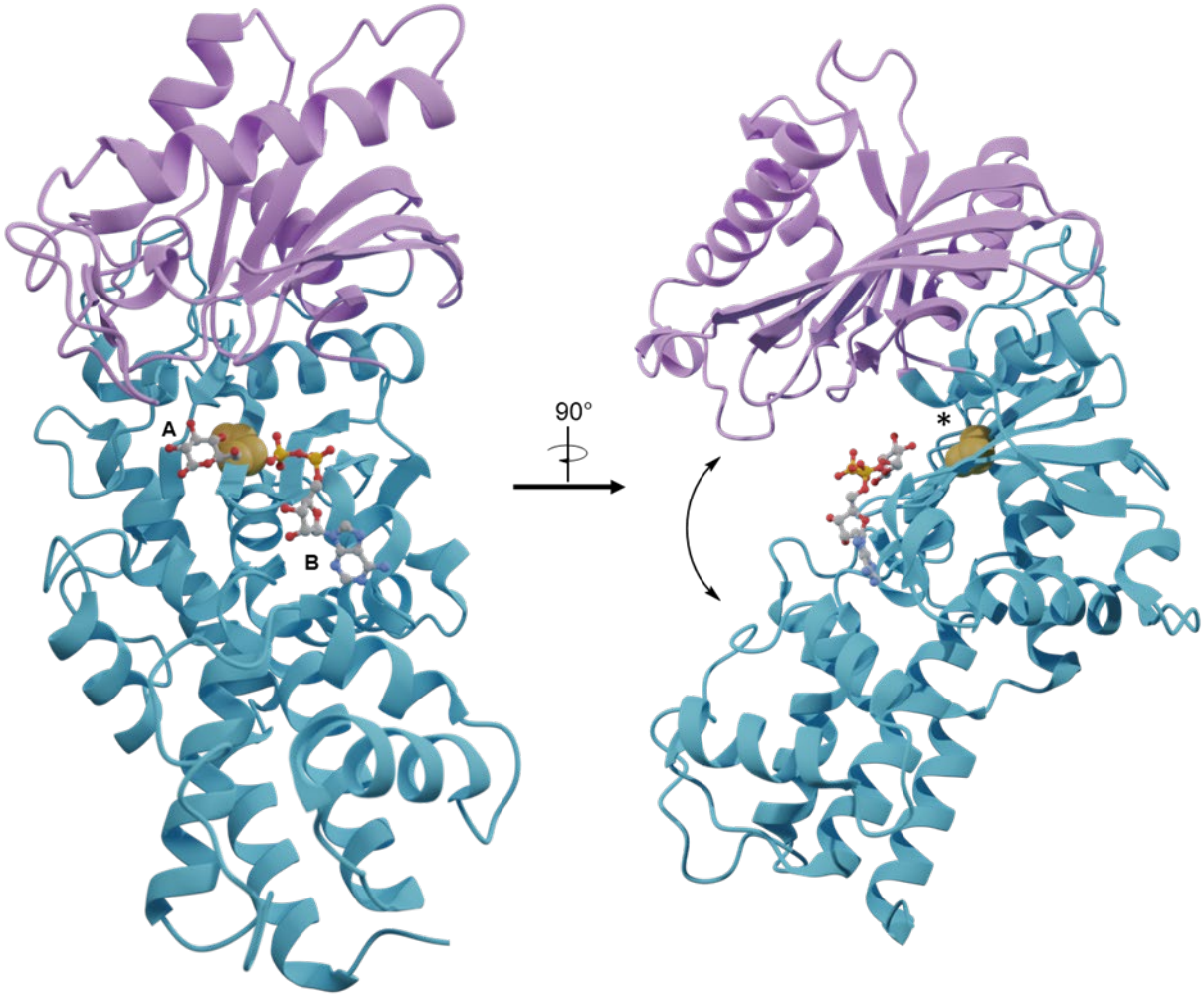


Figure 20. *Saccharomyces cerevisiae* hexokinase II structure and its G238V mutation (*). The model was generated by superimposing *Sc*Hxk2p (PDB 1IG8) with HsHk1 (PDB 4FPB) and OsHxk6 (PDB 6JJ8) to position glucose (A) and ADP (B), respectively. Shown in blue and pink are the large and small domains, respectively.

2-deoxy-glucose (2DG) is a toxic glucose analog missing the hydroxy group at the second position. 2DG, like glucose, is phosphorylated by hexokinases, producing 2DG-6P. 2DG-6P cannot undergo the next step in glucose catabolism, the isomerization to fructose-6-phosphate, because it is missing the required oxygen.[99] 2DG is a potent inhibitor of glycolysis even when other carbon sources are available. It inhibits cell growth by several mechanisms, including

weakening cell walls[100], repressing gene expression[101], and depleting ATP reserves[102], among others. Due to the increased energy demands in cancerous cells, 2DG has been extensively studied as a cancer therapeutic[103–105], but cancerous cells rapidly acquire resistance, undermining treatment.

To better understand 2DG resistance mechanisms, we used *in vivo* evolution and whole genome sequence analysis to identify spontaneous mutations that confer resistance to 2DG in baker's yeast. We found a novel mutation in hexokinase II that confers resistance to 2DG. This mutation does not line the enzymatic cleft, but it still affects the enzymatic activity. We performed four sets of molecular dynamics simulations to elucidate the mechanism by which this mutation confers 2DG resistance. These simulations suggest that the mutation alters the dynamics of the glucose binding cleft, discouraging persistent glucose binding. Our findings provide novel insight into the mechanisms that cancer cells use to acquire 2DG resistance, which will aid in designing more effective hexokinase inhibitors for cancer treatment.

3.2 Methods

3.2.1 Experimental section methods

I did not perform any of the physical experiments reported in the manuscript. For completeness' sake, I have copied them *verbatim* in the sections below.[90]

3.2.1.1 Yeast strains, plasmids, and growth conditions

Yeast strains employed in this study are listed in Table 3 (reproduced in Figure 21). The strains were grown on YPD (2% peptone, 1% yeast extract, 2% glucose) or synthetic complete medium (per O'Donnell *et al.*[106]) lacking the amino acids needed for maintaining plasmids. Plasmid information is provided in Table 4 (reproduced in Figure 22). Plasmids were introduced into yeast strains using the lithium acetate transformation method.[107] Where indicated, SC or YPD containing 2% glucose were supplemented with 2DG to a final concentration (presented as % w/v). We generated a 2% 2DG (Sigma-Aldrich, St. Louis MO) stock by dissolving two grams of 2DG in 100 mL of water and then filter sterilizing. Unless otherwise indicated, cells were grown at 30°C.

Strain	Genotype	Source
BY4742	MATα his3Δ1 leu2Δ0 lys2Δ0 ura3Δ0	[128]
Parental ABC16-monster (RY0568)	***MAT-α*** *adp1Δ snq2Δ ycf1Δ pdr15Δ yor1Δ, vmr1Δ pdr11Δ, nft1Δ bpt1Δ ybt1Δ ynr070wΔ yol075cΔ aus1Δ pdr5Δ pdr10Δ pdr12Δ can1Δ::GMToolkit-α* [CMVpr-rtTA NATMX4 STE3pr-LEU2] *his3Δ1 leu2Δ0 ura3Δ0 met15Δ0*	[45]
Naïve ABC16-monster	***MAT-α*** *adp1Δ snq2Δ ycf1Δ pdr15Δ yor1Δ, vmr1Δ pdr11Δ, nft1Δ bpt1Δ ybt1Δ ynr070wΔ yol075cΔ aus1Δ pdr5Δ pdr10Δ pdr12Δ can1Δ::GMToolkit-α* [CMVpr-rtTA NATMX4 STE3pr-LEU2] *his3Δ1 leu2Δ0 ura3Δ0 met15Δ0*	[45]
2DG Resistant Strain 1 (ABC16-monster)	***MAT-α*** *adp1Δ snq2Δ ycf1Δ pdr15Δ yor1Δ, vmr1Δ pdr11Δ, nft1Δ bpt1Δ ybt1Δ ynr070wΔ yol075cΔ aus1Δ pdr5Δ pdr10Δ pdr12Δ can1Δ::GMToolkit-α* [CMVpr-rtTA NATMX4 STE3pr-LEU2] *his3Δ1 leu2Δ0 ura3Δ0 met15Δ0, hxk2^{G238V}*	This study.
2DG Resistant Strain 2 (ABC16-monster)	***MAT-α*** *adp1Δ snq2Δ ycf1Δ pdr15Δ yor1Δ, vmr1Δ pdr11Δ, nft1Δ bpt1Δ ybt1Δ ynr070wΔ yol075cΔ aus1Δ pdr5Δ pdr10Δ pdr12Δ can1Δ::GMToolkit-α* [CMVpr-rtTA NATMX4 STE3pr-LEU2] *his3Δ1 leu2Δ0 ura3Δ0 met15Δ0, hxk2^{G238V}*	This study.
2DG Resistant Strain 3 (ABC16-monster)	***MAT-α*** *adp1Δ snq2Δ ycf1Δ pdr15Δ yor1Δ, vmr1Δ pdr11Δ, nft1Δ bpt1Δ ybt1Δ ynr070wΔ yol075cΔ aus1Δ pdr5Δ pdr10Δ pdr12Δ can1Δ::GMToolkit-α* [CMVpr-rtTA NATMX4 STE3pr-LEU2] *his3Δ1 leu2Δ0 ura3Δ0 met15Δ0, hxk2^{G238V}*	This study.
2DG Resistant Strain 4 (ABC16-monster)	***MAT-α*** *adp1Δ snq2Δ ycf1Δ pdr15Δ yor1Δ, vmr1Δ pdr11Δ, nft1Δ bpt1Δ ybt1Δ ynr070wΔ yol075cΔ aus1Δ pdr5Δ pdr10Δ pdr12Δ can1Δ::GMToolkit-α* [CMVpr-rtTA NATMX4 STE3pr-LEU2] *his3Δ1 leu2Δ0 ura3Δ0 met15Δ0, hxk2^{G238V}*	This study.
2DG Resistant Strain 5 (ABC16-monster)	***MAT-α*** *adp1Δ snq2Δ ycf1Δ pdr15Δ yor1Δ, vmr1Δ pdr11Δ, nft1Δ bpt1Δ ybt1Δ ynr070wΔ yol075cΔ aus1Δ pdr5Δ pdr10Δ pdr12Δ can1Δ::GMToolkit-α* [CMVpr-rtTA NATMX4 STE3pr-LEU2] *his3Δ1 leu2Δ0 ura3Δ0 met15Δ0, hxk2^{G238V}*	This study.
MSY1254	***MAT-α*** *ura3Δ0 leu2Δ0 his3Δ1 hxk2Δ::KANMX4*	[40]
MSY1475	***MAT-α*** *ura3Δ0 leu2Δ0 his3Δ1 met15Δ0 hxk1Δ::KANMX4 hxk2Δ::KANMX4 glk1Δ::KANMX4*	[40]

https://doi.org/10.1371/journal.pcbi.1009929.t003

Figure 21. Yeast strains used in the current study. Reproduction of table 3 of the manuscript.[90]

3.2.1.2 In vitro evolution and whole genome sequencing analysis

We used directed evolution to identify mutations that confer 2DG resistance to the ABC16-monster strain,[108–111] which lacks sixteen ABC transporters. We evolved resistance via serial passaging in five independent replicates. In each passage, cultures were grown at 30°C in 30 mL of YPD (2% peptone, 1% yeast extract, 2% glucose) and 2DG, with shaking at 250 rpm. We stopped each passage when the growth reached saturation (OD ~3.0 per visual inspection) and examined the cultures under a microscope to verify that there was no contamination. We then

placed 300 μL aliquots into a fresh supply of 30 mL YPD with 2DG (i.e., a 1:100 dilution into fresh media with drug) and repeated the process. In early passages, 0.05% 2DG was added, and growth to saturation required 4–5 days. As resistance developed, the time needed for saturation shortened to roughly two days. To ensure evolved resistance at higher 2DG concentrations, we then increased the 2DG concentration to 0.2% and resumed serial passages. Each replicate required between eight and twelve passages total, at which point the growth rate of each had stabilized (per eighteen-hour growth curves calculated at multiple 2DG concentrations, 0.05–0.2%). To enable comparative genomics and growth-rate analyses, we also generated a no-drug control that involved passaging for the same time intervals but in medium containing no 2DG.

Name	Description	Source
pRS313	CEN HIS3	[129]
pHxk2-3V5-313	Genomic clone of HXK2 with 592 bp upstream of ATG and 373 bp downstream of the stop and a C-terminal fusion to 3V5; CEN HIS3	This study.
pRS315	CEN LEU2	[129]
pRS315-Hxk2-GFP	Genomic clone of HXK2 with 592 bp upstream of ATG and 373 bp downstream of the stop and a C-terminal fusion to GFP; CEN HIS3	This study.
pRS315-Hxk2-G238V-GFP	Genomic clone of HXK2 with 592 bp upstream of ATG and 373 bp downstream of the stop and a C-terminal fusion to GFP; the G238V mutation was introduced by site-directed mutagenesis; CEN HIS3	This study.
pRS315-Hxk2-3V5	Genomic clone of HXK2 with 592 bp upstream of ATG and 373 bp downstream of the stop and a C-terminal fusion to 3V5; CEN HIS3	[40]
pRS315-Hxk2-G55V-3V5	Genomic clone of HXK2 with 592 bp upstream of ATG and 373 bp downstream of the stop and a C-terminal fusion to 3V5; the G55V mutation was introduced by site-directed mutagenesis; CEN HIS3	[40]
pRS315-Hxk2-G238V-3V5	Genomic clone of HXK2 with 592 bp upstream of ATG and 373 bp downstream of the stop and a C-terminal fusion to 3V5; the G238V mutation was introduced by site-directed mutagenesis; CEN HIS3	This study.

https://doi.org/10.1371/journal.pcbi.1009929.t004

Figure 22. Plasmid DNA used in the current study. Reproduction of table 3 of the manuscript.[90]

To determine the genomic changes associated with evolved 2DG resistance, we isolated the genomic DNA of both the resistant (passaged) and control strains using a glass-bead/phenol-extraction protocol.[112] We performed next generation sequencing on Illumina NextSeq500 machines. As detailed in Soncini *et al*.,[113] sequencing libraries were prepared and multiplexed into single lanes for each strain to produce 151-bp paired end reads.

To identify potential resistance-conferring mutations, we followed the protocol described in Ellison et al.[114] In brief, we used the Bowtie 2 software[115] to align the sequence reads to the

S288C reference yeast genome. We then used Samtools 1.3.1[116] to sort the alignments by their leftmost coordinates and to index the sorted alignments. BCFtools 1.3.1[117] was used for variant calling (consensus calling model). VCFtools 0.1.14[117] was used to identify variants that differed between the 2DG-resistant and no-drug control strains. Finally, we used SnpEff 4.3p[118] to annotate the identified variants (*e.g.*, frameshift variants, missense variants, stop-gained variants, disruptive inframe insertions, putative impact high/moderate/low, etc.).

3.2.1.3 2-deoxyglucose resistance assays

We monitored resistance to 2DG in three different ways. First, to verify the 2DG resistance of the passaged strains, we performed serial dilution growth assays by plating serial dilutions of yeast cells onto solid agar medium containing the indicated concentrations of 2DG and allowing cells to grow for the time indicated for each figure at 30℃. We compared the growth of the evolved yeast cells to the unpassaged, parental strain (ABC16-monster) or the parental strain that was passaged using medium lacking 2DG (naïve ABC16-monster). Serial dilution growth assays were performed as described in O'Donnell *et al.*[106] In brief, we grew cells to saturation overnight in YPD or SC medium, measured the optical density of each culture, and initiated our dilution series with a cell density of A600 = 1.0 (or ~1 x 107 cells/mL). We then made five-fold serial dilutions of cells and pinned them onto solid YPD or SC with or without 2DG (0.05%, 0.2%, and 0.4%).

Second, we assessed our 2DG-resistant or control strains (parental ABC16-monster or naïve ABC16-monster, as indicated above) using growth curve analyses.[119] In brief, we grew cells to saturation in YPD or SC medium, washed cells into fresh medium, and inoculated in triplicate into flat-bottom 96-well plates at an A600 of 0.05 in the medium indicated (i.e., YPD or SC containing varying concentrations of 2DG). Prepared plates were incubated with shaking in a BioTek Cytation 5 plate reader (BioTek instruments; Winooski, VT, USA), and optical density

measurements were taken every 30 minutes for 24 hours using the Gen5 software package. Optical densities measured over time are presented with a path-length correction (to report measurements in a 1 cm path length). We used these curves to calculate the doubling times of yeast cells via the following equation:

$$doubling\ time\ = \frac{\ln(2)}{\left(\frac{\ln(OD_2) - \ln(OD_1)}{t_2 - t_1}\right)}$$

Doubling times were calculated based on the mean growth curves of each strain by selecting two points that span the linear range of the logarithmic growth portion of the growth curve. Third, we challenged cells with a range of 2DG concentrations, as described in Soncini et al.[113] In this approach, overnight cultures are grown to saturation in either glucose (Fig 3A) or galactose (Fig 5D) as a carbon source, diluted to an A600 of 0.1, and grown in the absence or presence of 2DG (0.01%, 0.02%, 0.05%, 0.1%, or 0.2%) for 18 hours at 30℃ with either 2% glucose (Fig 3A) or 2% galactose (Fig 5D) as a carbon source.[113] Each A600 was measured, and cell growth was normalized to growth in the absence of 2DG for each strain. The average of three replicate cultures is presented in Fig 3A, with statistical comparisons made using the Student's t-test for unpaired variables with equal variance. In this case, p-values are indicated as follows: *p < 0.05, **p< 0.01, ***p < 0.001.

3.2.1.4 Immunoblotting to assess Hxk2G238V abundance and stability

To assess Hxk2G238V abundance in cells, we performed whole cell protein extracts using the trichloroacetic acid (TCA) method.[106,120] In brief, an equal density of mid-log phase cells was harvested by centrifugation, washed in water, and then resuspended in water with 0.25 M sodium hydroxide and 72 mM β-mercaptoethanol. Samples were then incubated on ice, and proteins were precipitated by the addition of TCA. After incubation on ice, proteins were collected as a pellet by

centrifugation, the supernatant was removed, and the proteins were solubilized in 50 µL of TCA sample buffer (40 mM Tris-Cl [pH 8.0], 0.1 mM EDTA, 8M urea, 5% SDS, 1% β-mercaptoethanol, and 0.01% bromophenol blue). Samples were then heated to 37℃ for 30 minutes, and the insoluble material was removed by centrifugation before resolving samples by SDS-PAGE. Proteins were transferred to a membrane support and detected with either anti-GFP antibodies (Santa Cruz Biotechnology) or an anti-V5 probe (Invitrogen), followed by goat anti-mouse IRDye 680 (Thermo) or goat anti-rabbit IRDye 800 (LiCor). Antibody complexes were visualized using an Odyssey Infrared Imager (LiCor), and bands were quantified using the Odyssey software. REVERT (LiCor) total protein staining of membranes was used as a protein loading and membrane transfer control in immunoblotting.

3.2.1.5 Enzymatic assays for Hxk2 function

To verify that the hxk2G238V mutation alone is sufficient to confer 2DG resistance, we used site-directed mutagenesis to introduce the hxk2G238V mutation into a plasmid encoding the HXK2 gene (Table 4). We performed DNA sequencing of the entire open reading frame to ensure that no unintentional changes were generated. We separately transformed plasmids expressing WT HXK2 and hxk2G238V into the hxk1Δ hxk2Δ glk1Δ triple deletion cells (Table 4) and measured the hexokinase activity associated with these two alleles, as described in Soncini et al.[113] In summary, we prepared protein extracts using a glass-bead extraction protocol and assayed enzymatic activity by coupling the phosphorylation of glucose to its oxidation by glucose-6-phosphate dehydrogenase. The resulting production of NADPH, detected by measuring absorbance at 340 nm, correlates with hexokinase activity. For comparison, we used the same protocol to assess the enzymatic activity of WT Hxk2 (positive control). To measure the Michaelis-Menten constant (Km), we measured the reaction rate (v) at several glucose or 2DG concentrations

([S]) using a constant concentration of ATP (1 mM) and plotted the inverse of rate (1/v) against the inverse of concentration (1/[S]) (Lineweaver-Burk plot).[113] To calculate the Km for ATP, we measured the reaction rate at several ATP concentrations, kept the glucose concentration constant (2 mM), and plotted the inverse rate against the inverse of the substrate concentration.

3.2.1.6 Invertase assays

The invertase activity of cells grown in 2% glucose, where expression of the SUC2 gene that encodes invertase is repressed in an Hxk2-dependent manner, was measured as in Soncini et al.[113] For this assay, three independent cultures were assessed using a colorimetric assay that measures Suc2 enzymatic function coupled to glucose oxidase.[121] The mean of these replicates is plotted with the standard error indicated by the error bars. Invertase activity is measured in units per OD of cells, where 1 unit is equal to 1 μmole of glucose released per minute. Student's t-test for unpaired variables with equal variance was used to compare the difference between hxk2Δ cells containing plasmids expressing WT HXK2 vs. an empty vector or the hxk2 mutant alleles. P-values are indicated as follows: *$p < 0.05$, **$p< 0.01$, ***$p < 0.001$.

3.2.2 File preparation

The crystal structure of the *apo Sc*Hxk2p (PDB 1IG8[52]) was downloaded from the Protein Data Bank. We obtained the mutated protein by computationally mutating G238 to valine. Jennifer Walker changed the experimentally observed mutation using the Mutation-Wizard tool in PyMOL.[122] For the protein bound to glucose (*holo*), I manually generated the model because no crystal structure of the complex exists. The structure of Hxk2 bound to a ligand with a glucose-like substructure (PDB 2YHX[123]) has the same sugar binding pose as *Sc*Hxk1p (yeast, PDB

3B8A[96]), *hs*Hk2 (human, PDB 2NZT[124]), and *hs*Hk1 (human, PDB 4FPB), which are all bound to glucose. Given this consistent glucose binding pose, I generated the *holo* model by superimposing the *apo* structure and the crystallographic structure of *Sc*HxkIp bound to glucose (PDB 3B8A).

I used the PDB2PQR[53] tool to protonate Hxk2 at a pH of 7. This program uses the PROPKA[54] algorithm to optimize the hydrogen-bond network. Using LEaP from Ambertools18[55], I added water molecules and additional Na^+ and Cl^- counterions to neutralize the environment and approximate a 150 mM solution. Water molecules extended 10 Å beyond the protein in all directions. I used the Amber's ff14sb[56], tip3p[57], and GLYCAM_06j-1[125] force fields for the protein, water, and glucose.

3.2.3 Molecular dynamics simulations

All molecular dynamics simulations were done with the NAMD 2.13 package.[58,126] We performed a stepwise minimization, each consisting of 5000 steps. First, we relaxed all hydrogens; then we added the water molecules to the atoms to be relaxed; followed by hydrogen atoms, water molecules, and protein side-chain relaxation; finally, all atoms were allowed to relax. The equilibration of the systems was done in the NPT ensemble (isobaric-isothermal) at 310K. We also equilibrated the *apo* simulations in a stepwise manner. Each step consisted of 0.25 ns, in which we applied constraints to the protein backbone atoms, starting at 1.0 kcal/mol/Å^2, which we gradually relaxed to 0.75, 0.5, 0.25, and finally 0.0 kcal/mol/Å^2. For the ligand-protein complex, I relaxed in a single unrestrained one ns equilibration step that used one fs timestep. All simulations used the SHAKE[127] algorithm, the Nosé-Hoover method to maintain a pressure of 1.01325 bar, and the Langevin thermostat (5 ps^{-1} collision frequency).

Following the minimization and equilibration of the four systems, we started three independent isothermal-isobaric (NPT) production runs per system. Two of the three production runs consist of 250 ns, while the other consists of 500 ns. Each simulation uses a two fs timestep.

3.2.4 Analysis of molecular dynamics simulations

To confirm that our simulations had equilibrated after the minimization and equilibration steps, I calculated the RMSD of the backbone heavy atoms against the first frame of the simulation using MDAnalysis 1.0.0.[60,61] Plots of the calculated RMSD values for the 12 simulations (divided into four systems) showed that the systems had not fully equilibrated. I considered the first five ns of the production runs as part of the equilibration step and discarded them. All further analysis does not include these discarded portions of the simulations.

RMSD against the first frame of the simulation and per-residue RMSF analyses were performed using MDAnalysis 1.0.0. Per-residue RMSF is a metric to assess the flexibility of each residue, and I used the center of geometry of each residue to calculate this metric. To evaluate the protein's opening and closing mechanism, a large-scale conformational change, I calculated the radius of gyration (RoG) of the protein in each frame of the simulation. RoG is a measure of protein conformation compactness. I used PDB 1IG8[52] and 3O8M[87] as the open- and close-conformation references. I confirmed the results obtained by RoG by monitoring the change in the distance between the centers of geometry of the big and small domains.

3.2.5 Dynamical cross-correlation (DCC)

I calculated DCC matrices for each of the four systems using MD-TASK.[128] Values in a DCC matrix describe the correlated motion between residue *i* and residue *j* (*i.e.*, 1 is a complete correlation, and -1 is a complete anticorrelation). Elements in this matrix are calculated as follows:

$$C_{ij} = \frac{\langle \Delta r_i \cdot \Delta r_j \rangle}{\sqrt{\langle \Delta r_i^2 \rangle} \cdot \sqrt{\langle \Delta r_j^2 \rangle}}$$

To observe changes in correlation between residues among systems, I calculated ΔDCC matrices by element-wise subtraction.

3.2.6 Betweenness centrality (BC)

I calculated BC using MD-TASK. BC is a metric that describes the importance of a node for communication within the network. To calculate BC, we represent the protein as a graph with nodes centered at the Cβ atoms of each residue (Cα for glycine) and edges connecting any two nodes within 6.7 Å of each other. With the graph representation of the protein, one can calculate the shortest paths between all residue pairs. The BC of a node (a residue in our case) is the number of shortest paths that pass through that node. BC is a metric of the importance a given node has for communication within the network. I normalized the per residue BC value in each of the systems.

3.3 Results and discussion

3.3.1 Summary of the experimental section

I am presenting the experimental results obtained mainly by Mitchell Lesko and Jennifer Walker to give context to the computational experiments and results. These experiments were not performed by me.

To identify new mechanisms of 2-deoxyglucose (2DG) resistance, we performed in vitro evolution and whole genome analysis.[129,130] We used an especially sensitive *Saccharomyces cerevisiae* (baker's yeast) strain that lacks 16 ABC transporters (ΔABC16).[131–134] These cells are less able to evade cytotoxic chemicals by an efflux mechanism. We exposed ΔABC16 to increasing concentrations of 2DG via serial passaging, resulting in five 2DG resistant strains.

We performed a whole genome analysis to find the genetic causes of resistance. All five evolved strains contained a missense mutation in Hxk2 at position 238. The WT protein has glycine, while the mutated protein has valine ($Hxk2^{G238V}$). To confirm that $Hxk2^{G238V}$ is enough to confer resistance, we introduced Hxk2, $Hxk2^{G238V}$, and $Hxk2^{D211A}$ to cells lacking Hxk2, which are inherently resistant to 2DG. We observed that both the catalytically inactive $Hxk2^{D211A}$ and the mutated $Hxk2^{G238V}$ were still resistant to 2DG, while the WT Hxk2, as expected, had restored sensitivity to 2DG. Cells that do not have Hxk2 are resistant to 2DG, which brings the possibility that the $Hxk2^{G238V}$ protein might be unstable. To assess its stability, we introduced $Hxk2^{G238V}$ into a cell line without any of the three hexokinases (*hxkIΔHxk2Δglk1Δ*). This line supported yeast growth on glucose, showing that $Hxk2^{G238V}$ is sufficiently folded to perform its enzymatic activity.

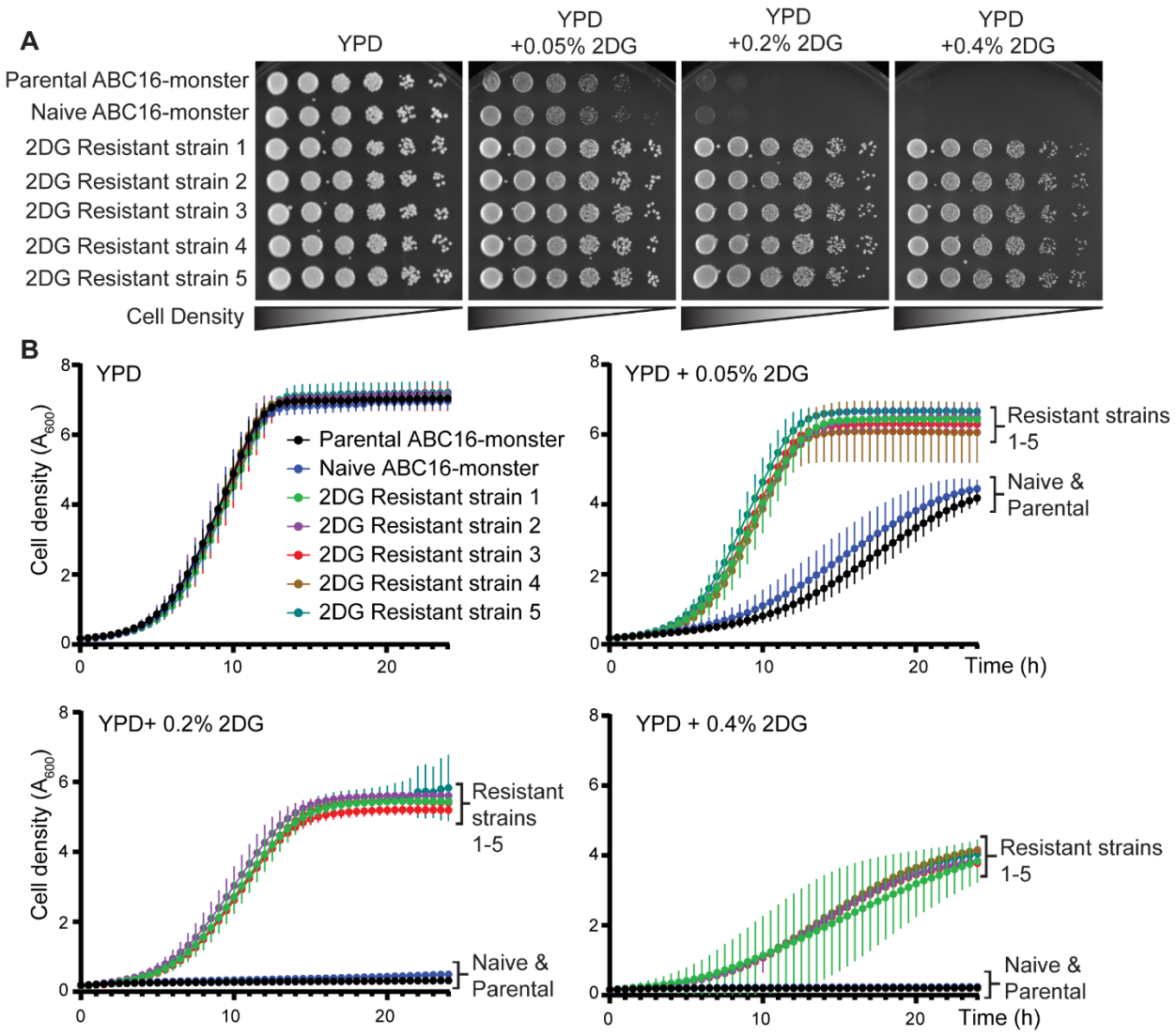


Figure 23. Strains 1-5 are resistant to 2DG. A) Images of serial dilution growth assays of parental ΔABC16, and five resistant strains. We used increasing concentrations of 2DG with 2% glucose as the carbon source. B) Plots showing the change in cell density over time. Image generated by Mitchell Lesko.

To assess changes in enzymatic activity in Hxk2^{G238V}, we compared the ability of Hxk2^{G238V} and Hxk2^{G238V} to phosphorylate glucose. We obtained protein extracts of *hxkIΔHxk2Δglk1Δ* cells after we had introduced either Hxk2 or Hxk2^{G238V}. With these extracts, we compared average NADPH production as a proxy of glucose-6-phosphate generation.[83] These

experiments revealed a dampened activity for Hxk2^{G238V}, with WT Hxk2 having a substantially higher specific activity and an appreciably lower K_m (see Table 1).

Table 1. Enzyme kinetics for WT Hxk2 and Hxk2^{G238V}. In parenthesis, we show statistical significance. * < 0.0005, ** < 0.005. Values of 2DG phosphorylation by mutant Hxk2 could not be reliably determined (ND) and are not shown. K_m represents the Michaelis-Menten constant, and SA is the specific activity (V_{max} normalized by the enzyme level).**

	K_m glucose (mM)	SA Glucose (nmol/min/au)	K_m 2DG (mM)	SA 2DG (nmol/min/au)	K_m ATP (mM)	SA ATP (nmol/min/au)
WT Hxk2	0.23 ± 0.02	27.6 ± 1.6	0.48 ± 0.06	8.59 ± 0.67	0.13 ± 0.01	24.2 ± 1.0
Hxk2^{G238V}	2.3 ± 0.37 (***)	8.9 ± 1.0 (**)	ND	ND	2.0 ± 0.21 (***)	7.1 ± 0.71 (**)

In summary, we demonstrate that the novel G238V mutation in Hxk2 is responsible for the observed 2DG resistance. Hxk2^{G238V} is a stable and functional protein, but the catalytic activity of this mutant is substantially lower than the WT protein.

3.3.2 The mutation is unlikely to interfere directly with ligand binding

To understand the molecular mechanism by which Hxk2^{G238V} elicits resistance to 2DG, I generated a glucose-bound model of Hxk2 (see Methods section). In this model, which was generated from the protein in the open conformation, I observed that the mutation is not likely to interfere directly with glucose binding. G238 is in β10, ~5.0 Å away from any glucose atom, and the mutated valine side chain would point towards the protein's interior, not towards bound glucose. To determine if the mutation would interfere with ligand binding or catalysis in the closed conformation, I used the glucose-bound *Kluyveromyces lactis* Hxk2 crystal structure as a model (PDB 3O8M[87]). I had to use this homologous protein (73.4% sequence identity with *Sc*Hxk2 per

Blast alignment[135,136]) because there is no ScHxk2 structure in the closed conformation. The distance between the G238 equivalent residue (also a glycine) and any glucose atom is 5.8 Å. This mutation is also unlikely to directly interfere with catalysis since it does not form any interactions with D211, the catalytic residue.

3.3.3 The Hxk2^{G238V} mutation may alter protein dynamics

To observe whether Hxk2^{G238V} changes protein dynamics in hopes of understanding the mechanism by which the mutation lowers enzymatic activity, I performed molecular dynamics of four different systems: *apo* WT protein (*apo* Hxk2), glucose bound WT protein (*holo* Hxk2), *apo* mutant protein (*apo* Hxk2^{G238V}), and glucose bound mutant protein (*holo* Hxk2^{G238V}). For each system, we performed three separate MD simulations, one of 500 ns and two of 250 ns, for a total of one μs per system (4 μs total). All systems started from the open conformation (PDB 1IG8). My goal with these glucose simulations was to capture the initial dynamics following glucose binding.

3.3.4 Pocket dynamics is affected by the Hxk2^{G238V} mutation

The simulations showed that three pocket regions were mainly affected by the mutation. The β9/β10 β-hairpin (I231-V236), the catalytic residue D211, and the α11 helix (D417-P425). The flexibility of these three regions is increased in the *apo* Hxk2^{G238V} simulations. In the *holo* simulations, the β-hairpin is stabilized by the mutation.

3.3.4.1 β9/β10 β-hairpin

The β9/β10 β-hairpin (I231-V236) is a protein region at the center of the enzymatic cleft, close to the ATP and glucose binding sites. This β-hairpin likely influences catalysis by its proximity to D211, the catalytic residue. To assess flexibility differences due to the mutation, I calculated the RMSF of each residue and its differences between simulations. These ΔRMSF ($RMSF_{WT} - RMSF_{G238V}$) calculations suggest that in the *apo* simulations, the mutation increases the flexibility over the WT; on the other hand, in the *holo* simulations, it is decreased compared to the WT.

The mutation affects the dynamics of V236, a β-hairpin residue. To monitor this change, I calculated the V236 χ_1 dihedral angle (N-Cα-Cβ-$C\gamma_1$) throughout the simulation (Figure 24C). For the Hxk2 simulations, we observe a change in V236 side-chain conformations due to glucose binding. In the *apo* state, Hxk2 heavily samples the *gauche* conformation, the same conformation observed in the open reference (-53.1°). However, in the *holo* form, V236 χ_1 shifts more towards the *anti*-conformation, observed in the closed state (155.3° observed in PDB 3O8M). This shift of dihedral angles due to ligand binding is not observed for the mutated protein. Another difference in $Hxk2^{G238V}$ is that it barely samples the other *gauche* conformation. I hypothesize that V236's rotational transition from the *gauche* to the *anti*-conformer are crucial for domain closure, and the changes observed in $Hxk2^{G238V}$ affect this process (see Figure 24).

I also observed changes in the correlated motions between residue 238 and residues in the β-hairpin, as observed using dynamical cross-correlation (DCC) analysis. Differences in DCC ($\Delta DCC = DCC_{WT} - DCC_{G238V}$) show more correlation in the $Hxk2^{G238V}$ simulations for this region, suggesting an allosteric influence due to the mutation. Many residues in the β-hairpin have ΔDCC values two standard deviations larger than the mean ΔDCC across all residues.

Table 2. Dynamical cross-correlation of residue 238 (G for WT or V for mutant) with β9/β10 β-hairpin residues.

resid	residue	DCC *apo* wt all	DCC *apo* mut all	DCC *holo* wt all	DCC *holo* mut all
231	I	0.65	0.69	0.64	0.68
232	F	0.50	0.55	0.58	0.51
233	G	0.28	0.31	0.21	0.32
234	T	0.18	0.32	0.04	0.33
235	G	0.10	0.38	0.08	0.36
236	V	0.21	0.62	0.35	0.56

3.3.4.2 Aspartate 211, the catalytic residue

The catalytic residue D211 also is affected by the G238V mutation. I only observed a difference between WT and mutant when no ligand was present; the mutation increased the residue's flexibility. To understand the impact of the mutation on D211 dihedral angles, I calculated Janin plots for the four systems. Figure 24A shows that the impact of the mutation on the *apo* simulations is negligible, but the effect is quite noticeable in the *holo* simulations. Here we observe a noticeable population shift, and the structural difference is the position of the carboxy group, shown in Figure 24B. This change in population for the ligand bound mutant, may position D211 in a non-favorable way to catalyze the reaction with glucose, lowering the activity of this enzyme.

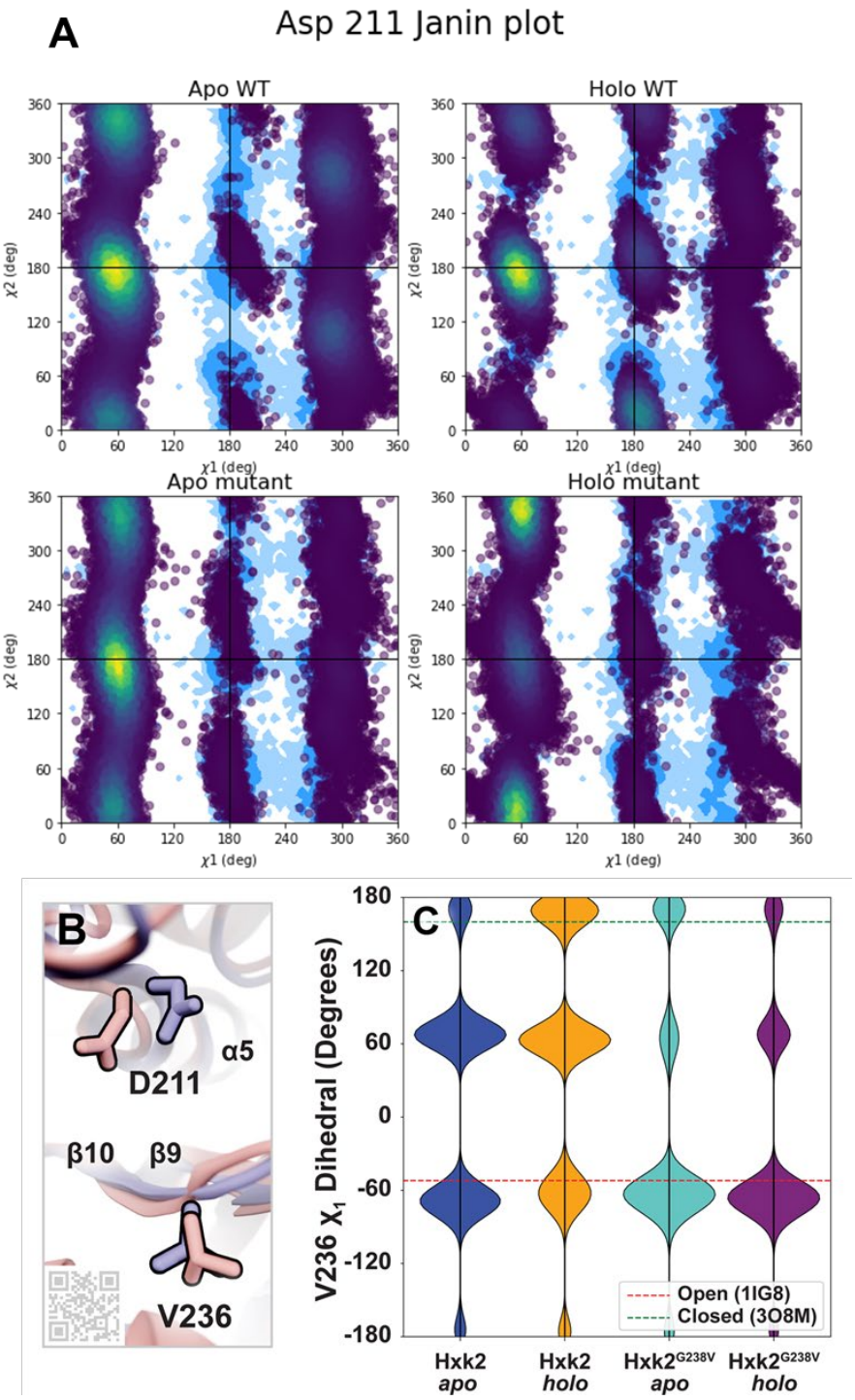


Figure 24. Dihedral angle analysis for D211 and V236. A) Janin plots for the catalytic aspartate (D211). B) Two conformations were obtained from the *apo* Hxk2^{G238V} simulations that show the conformational changes observed in D211 and V236. In blue, V236 is shown in an open-like confirmation, with D211 in a glucose-ready conformation. In pink, D211 and V236 are shown in displaced and closed-like conformations,

respectively. C) χ_1 dihedral angle distributions for V236. The angle typical of the open and closed conformations (references) are shown as dotted lines. Jacob D. Durrant made panel B.

3.3.4.3 α-helix 11

The third region affected by the mutation is α-helix 11 (D417-P425), a part of the protein lining the ATP-binding site. I observed increased flexibility for this helix in the Hxk2^{G238V} simulations only when no ligand was present. In contrast, I saw little to no effect in the ligand-bound simulations. The increased flexibility could suggest an impact on ATP binding, changing the catalytic properties of Hxk2.

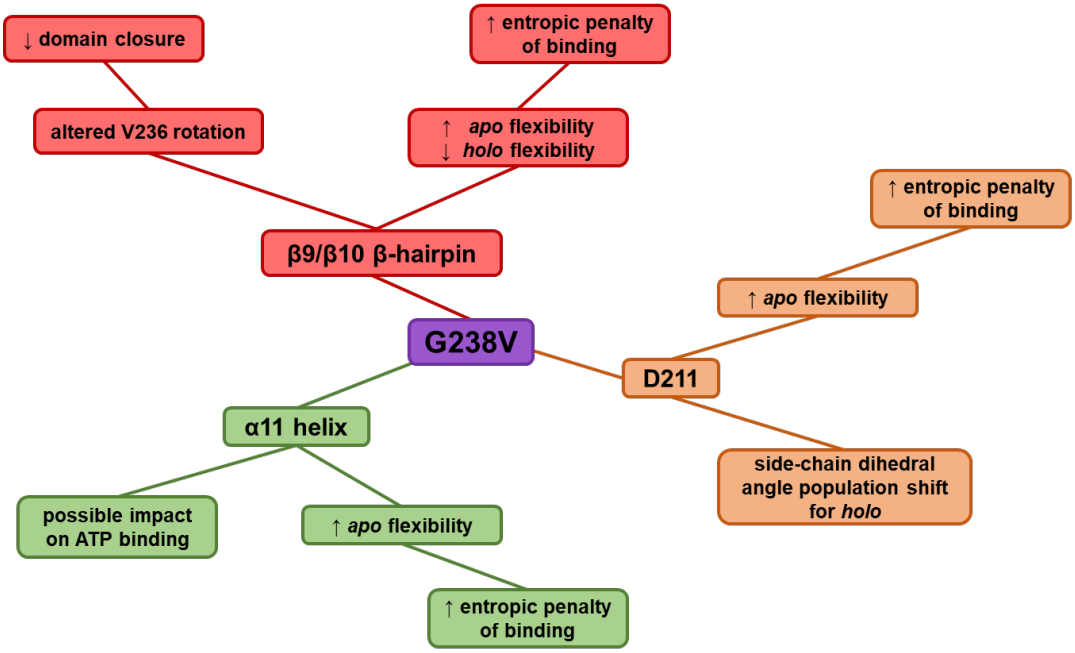


Figure 25. Summary of hypothesized impacts on Hxk2 pocket dynamics due to the G238V mutation. G238V affects three regions involved in Hxk2 catalysis: the β9/β10 β-hairpin, D211, and α11.

3.3.5 The Hxk2^{G238V} mutation affects global dynamics

To compare the simulated conformations to the open and closed conformations, I calculated the RMSD of each simulation against two references. For the open conformation, I used the backbone atoms of the *Sc*Hxk2p structure I used for the simulations, PDB ID 1IG8. For the closed conformation, I used a structure of *K. lactis* HxkIp (PDB 3O8M), as mentioned before. Although not a perfect match, *Kl*HxkIp structure helped us better understand the protein's opening and closing dynamics. The *apo* systems came close to the 1IG8 crystallographic structure, within 0.78 Å (RMSD) for the WT simulation and 0.75 Å for the mutant simulation. The *holo* simulations also came close to their closed-conformation reference (*Kl*HxkIp), within 1.13 Å and 1.18 Å for the WT and mutant simulations, respectively.

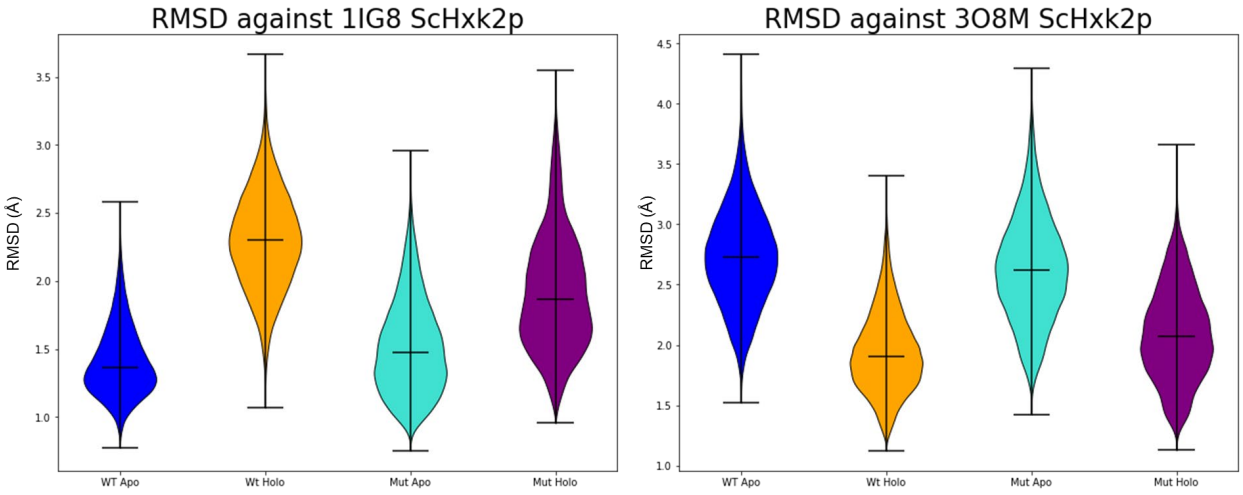


Figure 26. Violin plots showing the RMSD of the four systems compared to the open (PDB 1IG8) and closed (PDB 3O8M) reference conformations.

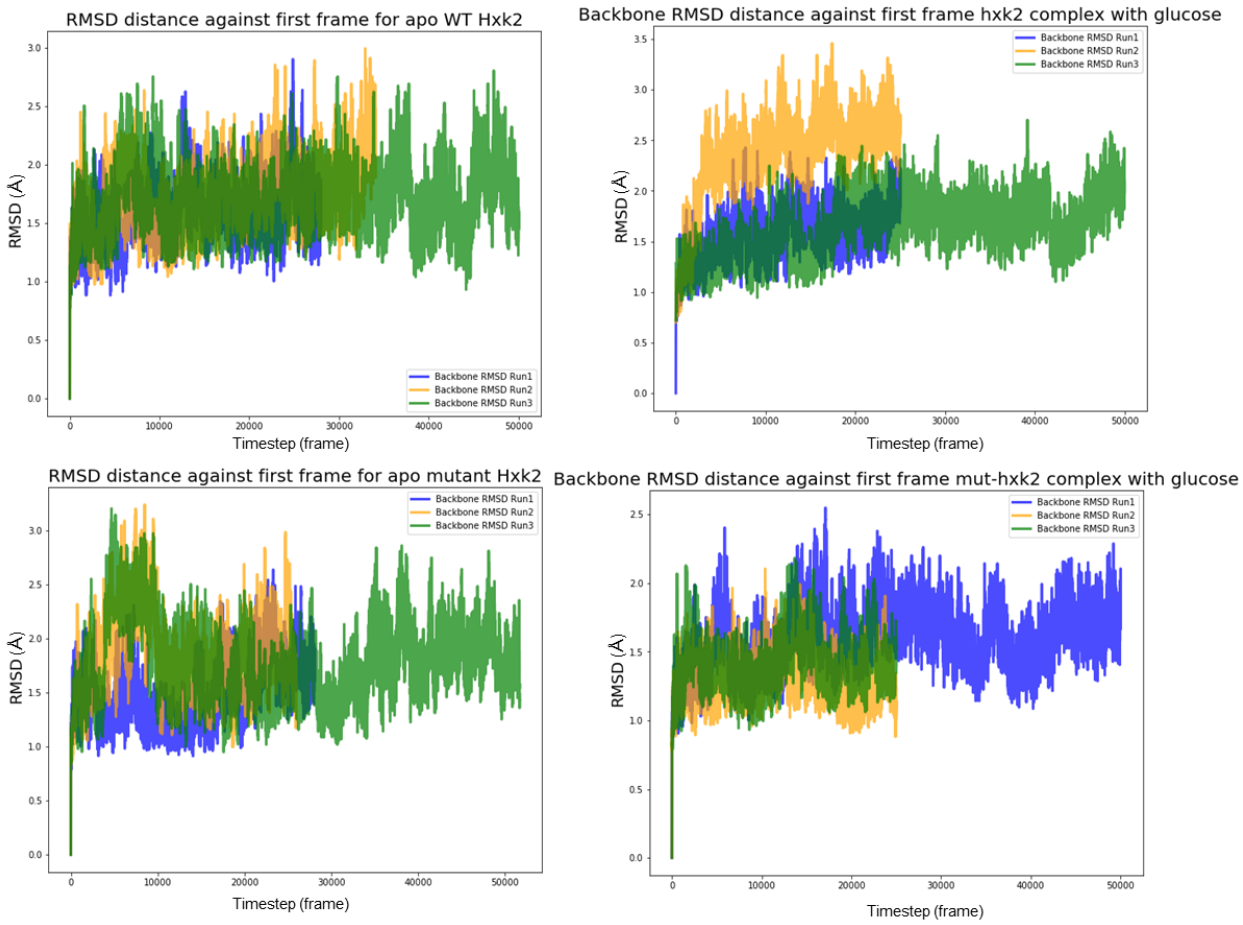


Figure 27. The RMSD between simulated conformations and the last frame of the equilibration simulation.

As an indirect way of measuring the opening and closing of the protein, I calculated the radius of gyration (RoG) of the whole protein using MDAnalysis. A larger RoG value indicates the protein is in the open conformation, and a lower value indicates the closed conformation. I performed a Kruskal-Wallis test to assess the difference in means because the RoG distributions are not normally distributed. I rejected the null hypothesis that the means of each system come from the same distribution (F-statistic of 2.6 x 10^6 and p-value < 0.001). I followed this analysis with the Conover *post hoc* analysis showing that all four systems are statistically different in terms of RoG. The RoG for Hxk2G238V is generally lower than for the WT protein, suggesting that Hxk2^{G238V} is less likely to adopt a fully open conformation. The mutation has a medium effect on RoG according to Cohen's d statistic (0.55 and 0.52 for *apo* and *holo* simulations, respectively).

The standard deviation is also different between the WT and Hxk2^{G238V}; the mutant protein has a greater standard deviation, suggesting it is less likely to adopt a fully closed conformation, perhaps explaining why its K_m is ten times higher than the WT.

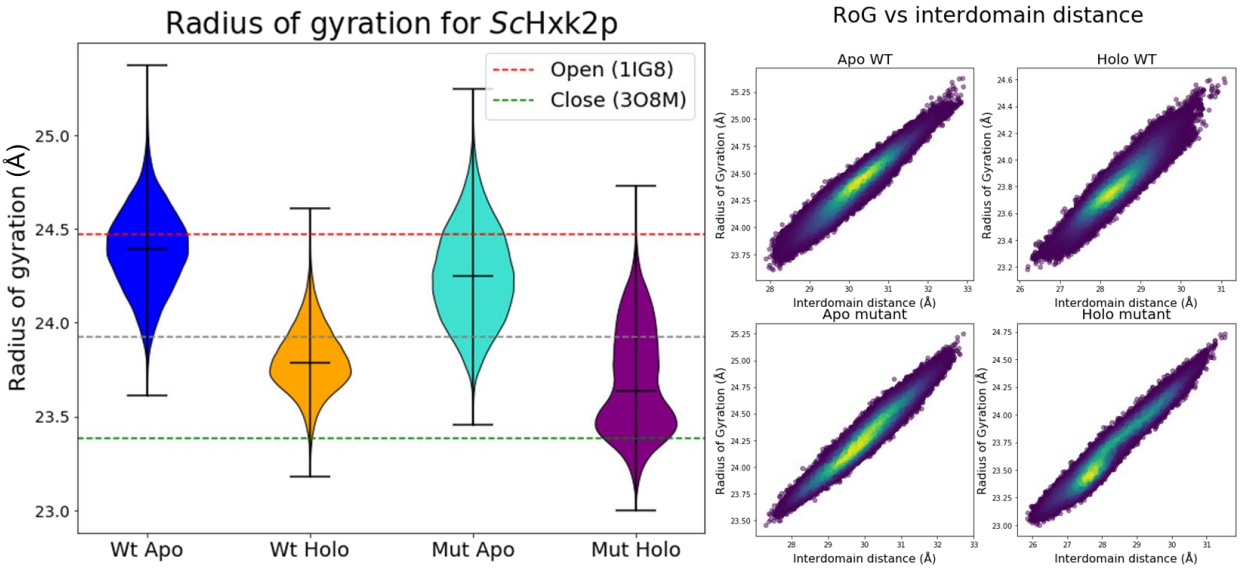


Figure 28. Opening and closing of Hxk2 as measured by the radius of gyration and the interdomain distance.

To corroborate the results obtained by the RoG analysis, I calculated the distance between the centers of geometry of the domains (Figure 28). I observe a linear correlation between the RoG and the interdomain distance, showing that RoG is a reasonable metric for how closed or open the protein is.

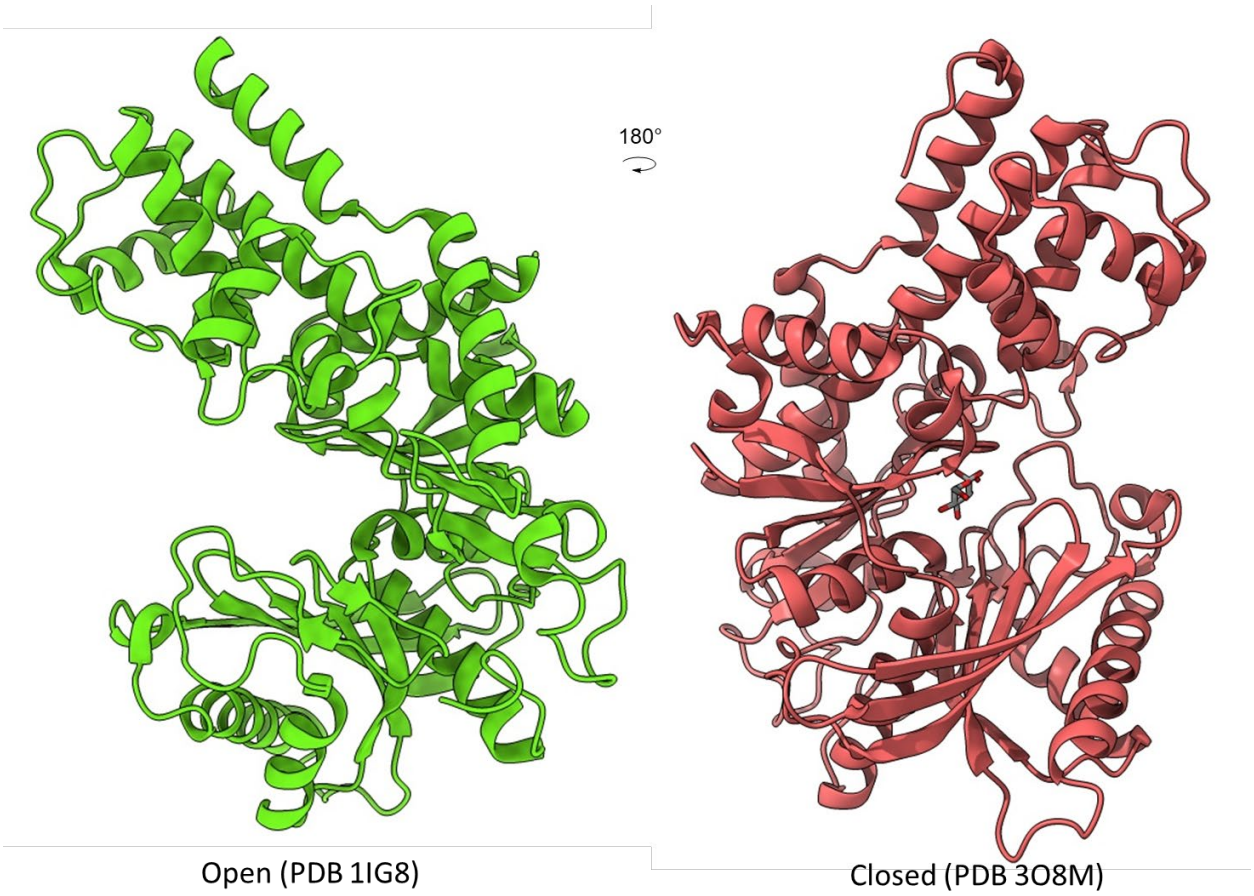


Figure 29. Comparison of the open and closed Hxk2 conformations using PDB 1IG8 and PDB 3O8M (*K. Lactis*)

3.3.6 Hxk2^{G238V} changes the centrality of cleft-lining residues

To assess changes in intra-protein communication, I calculated betweenness centrality (BC). In network analysis, BC measures the importance of a node (residue) in the flow of information within a network (protein). BC analysis showed how important D211 and the β-hairpin are for intra-protein communication (Figure 30). These residues are among the most connected residues, according to BC. These residues appear among the 90th percentile for at least one of the

simulated systems. All but T234 and G235 appear in the 90th percentile for all four systems. This suggests the β-hairpin is a nexus where information can flow.

The difference in the mutated residue itself is particularly noteworthy; in the WT simulations (glycine), this residue is in the 57th and 60th percentile for the *apo* and the *holo* simulations, respectively. But when residue 238 is a valine, it is among the residues in the 90th percentile for both simulations (*apo* 93rd and *holo* 94th percentile). The effect size observed for this change is large, with Cohen's d values of 2.53 for the BC difference in the *apo* simulations and 2.33 for the difference in the *holo* simulations (Hxk2 – Hxk2^{G238V}). This difference could be explained by the fact that the mutation exchanges a hydrogen atom for an isopropyl group. Increasing the volume of the side chain increases the possible interactions it can have with neighboring residues.

I calculated the per residue difference (Hxk2 – Hxk2^{G238V}) of BC values for the *apo* and *holo* simulations to understand how the flow of signals is affected due to the mutation. In the β-hairpin, I observed that all residues except T234 have at least one difference that is significant (two standard deviations above the mean). Especially V236, the BC value of this residue increases in both the *apo* and *holo* simulations by 145% and 133.5%, respectively. The increase in V236 for the mutant protein, suggests it becomes more critical for intra-protein communication in the mutant protein. The effect size observed for this change is large for both the *apo* and *holo* simulations, with Cohen's d values of 0.80 and 1.08, respectively.

The D211 BC values change significantly only in the *apo* simulations, with a small effect size (Cohen's d of 0.40). This residue is more interconnected in the WT protein. The changes in α11 are more varied, with some residues increasing BC values and others decreasing due to the mutation.

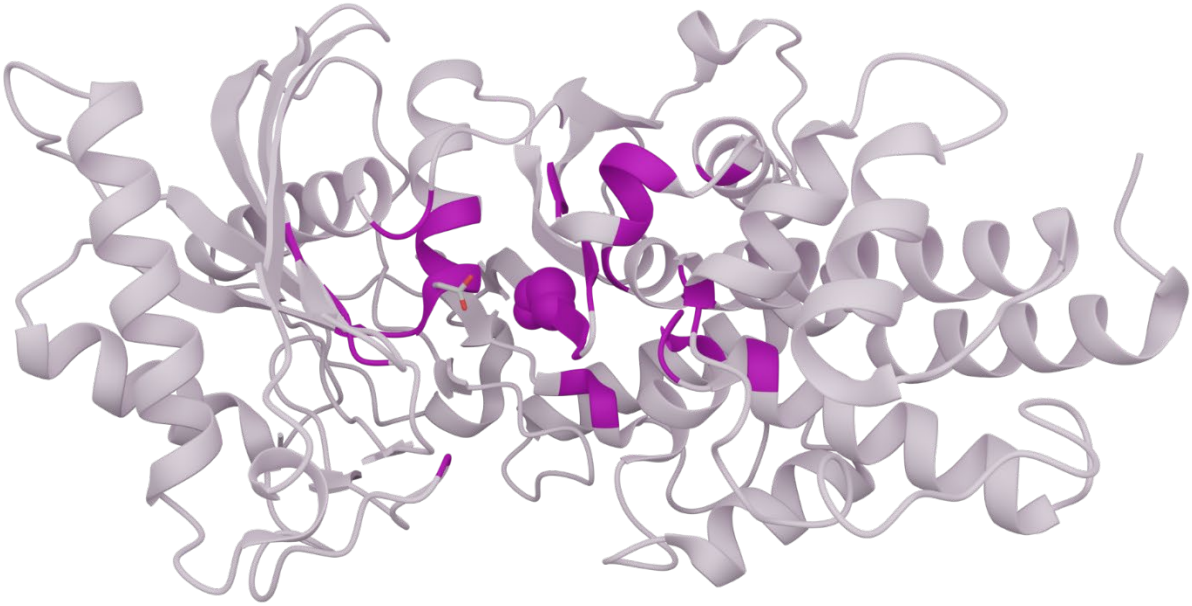


Figure 30. WT Hxk2. Residues in purple have significant changes in BC due to the mutation.

BC analysis suggests that the β-hairpin is a hub that communication can flow through, so any changes in its dynamics can be felt throughout the protein. I hypothesize that the binding of glucose could send a signal throughout the protein, inducing domain closure and catalysis. The shift of centrality to V236 in Hxk2^{G238V} may alter how the glucose binding signal propagates, impending domain closure, and catalysis.

3.4 Conclusions

In this work, we discovered a novel mutation in *Saccharomyces cerevisiae*'s hexokinase II that confers resistance to the known inhibitor 2-deoxy-glucose. Even though the catalytic activity of this mutant is substantially lower than the wild-type, it still provides the essential activity needed for survival. We demonstrated that the G238V mutation alone confers resistance to 2DG.

The addition of three heavy atoms due to the mutation appears to be a small change, but it has a crucial impact on the protein's dynamics. These changes, as I demonstrated, must be allosteric. In the mutant, V238 does not interact with glucose, but it changes the motions of the neighboring β-hairpin, which in turn propagates the signal to the rest of the protein, perturbing local and global dynamics. Primarily, we observed a substantial change in the dihedral angles sampled by V236, a residue that belongs to the β-hairpin. Its side chain undergoes a conformational change from *gauche* to *anti* when binding glucose in the WT simulations. This conformational change is also observed in crystallographic structures of the WT protein, but it does not occur in our simulations of the mutant protein. I hypothesize that the lack of rotational transition in V236 prevents the protein from sampling the full motion from the open to closed conformation. This in turn impacts glucose binding and catalysis.

The enzymatic-cleft flexibility is increased in $Hxk2^{G238V}$, which in turn increases the number of microstates visited by the protein; this increase may the entropic penalty of glucose binding for the mutated protein. Experimental methods like isothermal titration calorimetry (ITC), which can assess entropic and enthalpic contributions, could further explore this hypothesis.

This work is significant because, to our knowledge, this is the one of two 2DG resistance-conferring mutation observed in yeast's hexokinase II that does not directly impact the binding cleft. Most other mutations, directly impact glucose or ATP binding. (*e.g.,* glucose binding: K176T, T212P, Q299H; ATP binding: D211A, D417G, G418C, R423T, and S345P).[101,137]

The human homolog of this protein, *hs*Hxk2, is a potential target in cancer therapy. *hs*Hxk2 is usually upregulated in cancerous cells. These cancerous cells obtain their energy requirements through glycolysis and lactic acid fermentation, even in the presence of oxygen. 2DG binds *hs*Hxk2 and has anti-cancer properties, but spontaneous resistance has prevented its use in the

clinic.[138,139] Understanding the underpinnings of 2DG resistance is critical for developing new cancer therapies, and the present work provides insight into potential resistance mechanisms.

4.0 Mentoring undergraduates in computational research

4.1 Undergraduate research experiences, an introduction

Undergraduate research experiences (UREs) are experiences of students that join a faculty research laboratory, usually for longer than a semester. Regularly, they are provided with one-on-one mentoring, suggesting an apprenticeship model. We should not confuse UREs with course-based undergraduate research experiences (CUREs), which, as the name suggests, are part of a course, have a curriculum, and are broadly available to many students. CUREs usually replace a typical laboratory course.[140]

UREs' benefits range from increasing students' understanding of their particular field to developing practical skills like communication, synthesis of information, and problem-solving.[141] The main advantage of UREs is the retention of students in science, technology, engineering, and mathematics (STEM) majors.[142] UREs may also improve interest in higher degrees,[143] rates of acceptance to medical school,[142] development of self-identity as a scientist, and understanding of science practices,[140] among other benefits.[144] On the other hand, a study at a Hispanic serving institution showed that although Latino/a students are aware of such experiences, that awareness did not translate to engagement in UREs. Highlighting the need to fight preconceptions that underrepresented minorities might have about who can benefit from these experiences.[141]

Students participating in UREs are usually mentored by faculty members, postdoctoral researchers, or graduate students. The mentor's objective is to orient the mentee in developing and integrating concepts and background information as well as the technical aspects of the research.

But mentors also have the responsibility to teach the nature of scientific research and develop the mentee's science identity.[144]

4.2 Finding SMUG1 inhibitors, a traditional computer-aided drug design project

4.2.1 Justification

This project started as a summer research experience for Badiallo Diani, an undergraduate student, and is a collaboration with Professor Van Houten, who works at the Hillman Cancer Center. Ms. Diani was part of the TECBio 2021 initiative at the University of Pittsburgh. Due to the COVID restrictions, she thought this research experience would be an excellent opportunity to learn about the computational skills needed in computational biophysics research. The project aimed to find small-molecule binders of the target protein SMUG1, a base excision repair mechanism component. The student had to create a homology model of the protein and then dock a library of small molecules into the protein. I guided the student, met with her weekly, and was available through email to answer questions or schedule additional meetings if needed. I started by teaching the student the basics of the Linux terminal, reviewing the relevant techniques, and guiding her through the literature necessary to understand the methods she was using. Badiallo finished with a poster, and she found through docking a list of small-molecule candidate binders for the target protein.

After the summer, another undergraduate student, Ann Wang, completed the project. She had more experience with computational tools but was equally inexperienced in computational biophysics. This new student created a consensus list of the best docked small molecules and

visually inspected the best docking compounds. Using a more accurate docking algorithm, we redocked the best 250 small molecules according to the consensus list into the protein conformations.

4.2.2 Introduction

Deoxy ribonucleic acid (DNA), the molecule in charge of storing the genetic information within the cell, can be damaged at any time by several means (*e.g.*, radiation, reactive oxygen species, etc.).[145] This damage is highly relevant to cancer progression, its origins, and the advancement of tumors from benign to malignant. Cells have evolved several repair pathways to combat DNA damage, but cancerous cells often have weakened or defective repair mechanisms.[146,147] The base excision repair (BER) pathway is used to repair damage when it is localized in the bases. This damage can be due to oxidation, deamination, and alkylation, among others.[148] There are several proteins involved in BER, depending on the specific type of damage. Single-strand selective monofunctional uracil DNA glycosylase 1 (SMUG1), a member of the uracil-DNA glycosylases (UDG), is the protein responsible for the removal of oxidized pyrimidine, 5-hydroxymethyl-2'-deoxyuridine(5-hmdU), or the presence of uracil (due to the deamination of cytosine).[149] The name suggests SMUG1 acts only on single-stranded DNA, but this name is misleading since SMUG1 can act on double-stranded DNA. The only requirement for SMUG1 action on double-stranded DNA is for it to be potentiated by AP-endonuclease (APE1).[150] Besides maintaining DNA integrity, SMUG1 is involved in RNA maturation and quality control, as well as telomere maintenance.[149]

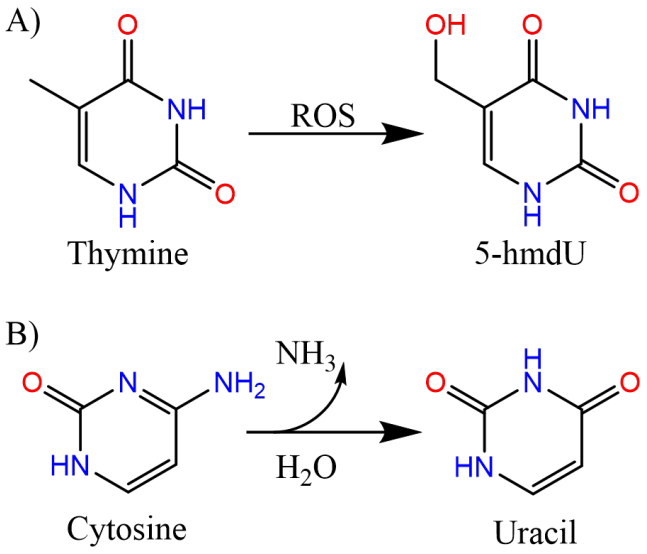


Figure 31. Lesions repaired by SMUG1. A) Formation of 5-hydroxymethyl-2'-deoxyuridine (5-hmdU) by reactive oxygen species. B) Formation of uracil in DNA.

SMUG1, as a member of the uracil DNA glycosylase (UDG) family, has the α/β fold characteristic of the UDG family, as seen by the crystallographic structure of *Xenopus* SMUG1 (xSMUG1 PDB 1OE4) bound to DNA.[151] Structurally, xSMUG1 has two unique features compared to other UDG members. These unique features help SMUG1 recognize damage in DNA. The first is a five amino acid loop (residues P251-P256, shown in orange in Figure 32), and the second is a five-residue long α-helix (residues P256-K260, shown in dark orange in Figure 32). xSMUG1's structure suggests it recognizes its substrate via a water displacement mechanism. Mayumi Matsubara and coworkers performed mutagenesis studies on human SMUG1 to identify critical residues for protein function. The catalytically relevant residues are N85 and H239. The

authors found that residues F98 and N163 are important for discriminating the pyrimidine rings, while residues G87, F89, G90, and M91 are critical for recognizing C5 substituents.[152]

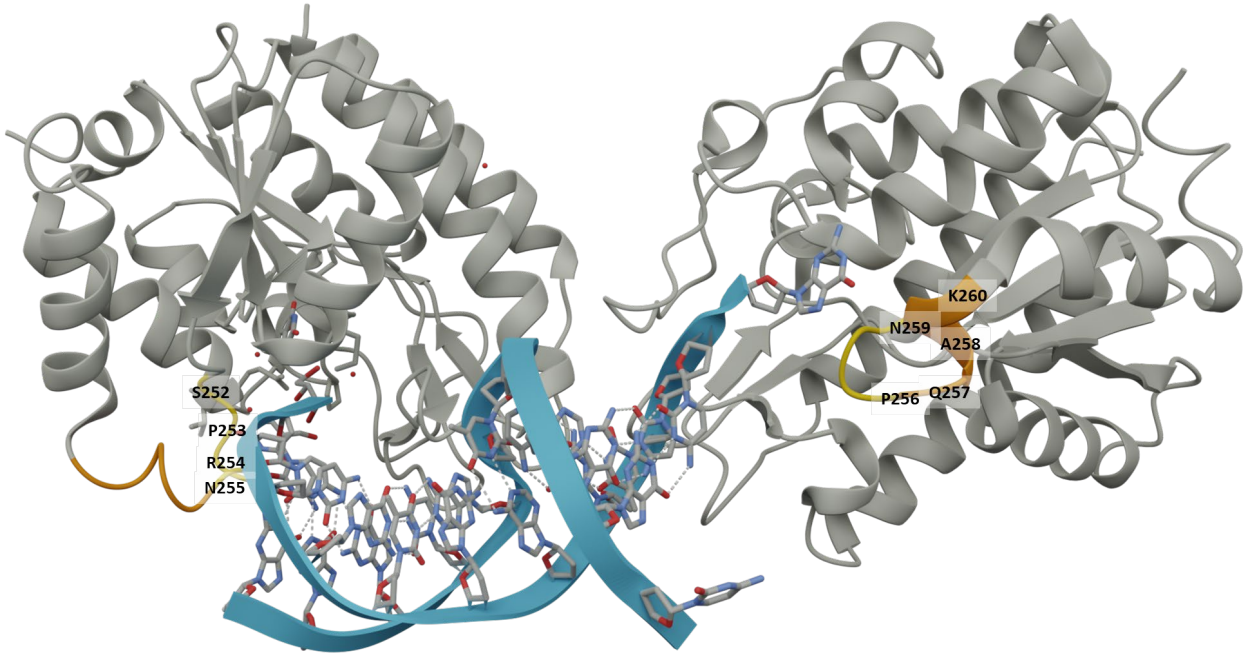


Figure 32. Crystal structure of Xenopus SMUG1 bound to DNA and soaked with 5-hmdU (PDB 1OE6). The unique loop and helix are shown in orange and red-orange, respectively.

SMUG1 plays a role in acquired resistance to 5-fluorouracil (FU), one of the most used drugs in chemotherapy. FUs cytotoxicity comes from its accumulation in the genome of the cancerous cell, and SMUG1 protects the cell by excising FU.[153] Damage by 5-hmdU, the primary substrate of SMUG1, is elevated in tumor cancer cells.[149] Finally, the action of FU was enhanced in the presence of 5-hmdU.[154] This suggests that SMUG1 could be a good drug target for novel cancer treatments.

We used several computational tools to identify probable hSMUG1 binders. To achieve the discovery of new drugs targeting hSMUG1, we modeled the hSMUG1 structure, obtaining six different conformations of the protein. Then with virtual screening techniques, we predicted the binding affinity of 13,421 small molecules. We performed docking on the six hSMUG1

conformations to account for protein flexibility, a protocol known as ensemble docking. It has been demonstrated that the use of different docking programs gives different results, and combining them in a technique called consensus scoring, can yield improved hit rates.[155,156] Using two different computer programs, we docked the small molecules into the proteins and created a consensus list with the best 250 predicted binders. These best binders were re-docked with a more accurate but computationally expensive algorithm to obtain 12 compounds with the best predicted binding affinities. Among these compounds, those that formed a salt bridge with residues R124, E135, or R243 tended to have lower predicted binding affinities (*i.e.*, they are better at binding hSMUG1). We sent the list of 12 molecules to our collaborators to test experimentally for hSMUG1 inhibition.

4.2.3 Methods

4.2.3.1 Homology model building

The sequence of human SMUG1 (hSMUG1) was submitted to the SWISS-MODEL web server to obtain a homo-dimer homology model (accessed on 2022-06-01).[157] This homo-dimer was then refined with DeepRefiner, a web server for high-accuracy protein refinement that uses neural networks (accessed on 2022-06-07).[158] We refined the protein as a dimer and monomer, using chain A as the input for the monomer algorithm. We used the default parameters for the refinement. Finally, we also used an AlphaFold structure of hSMUG1 (obtained on 2022-07-23).[14] This procedure gave us six different conformations for hSMUG1: SWISS-MODEL chain A (SM A), SWISS-MODEL chain B (SM B), DeepRefiner chain A (DR A), DeepRefiner chain B (DR B), DeepRefiner monomer (DR M), and AlphaFold (AF). The proteins were prepared for docking

with two packages: MGLTools and Maestro, for their use in Autodock Vina and Glide, respectively. We selected the orthosteric pocket and used the default parameters for both.

4.2.3.2 Small molecule datasets and ensemble docking

We considered three compound libraries: the NCI diversity set III, NCI diversity set VI, and the Enamine DDS-10-25-Y-10 set, with 1597, 1584, and 10240 compounds, respectively (13,421 small molecules total). These datasets were prepared in two separate ways. First, we generated three-dimensional structures using Gypsum,[159] followed by parametrization using MGLTools (Version 1.5.6).[160] MGLTools enables docking with Autodock Vina;[161,162] it calculates the partial charges of each atom, defines rotatable bonds, and adds polar hydrogens. Second, we used LigPrep,[163] a program in the Maestro suite (Version 2021-3),[164] a software package made by Schrödinger. We used default parameters, except that we constrained the generation of stereoisomers to a maximum of 16.

Initial docking of the 13,421 compounds into the six different conformations was performed using two different programs: Autodock Vina[161,162] and Glide[165166] (Schrödinger). The box size and center were manually selected, targeting the catalytic pocket. For Autodock Vina, we used a box size of 15 Å and exhaustiveness of 20. For Glide, we used the standard precision (SP) level of theory and the default parameters.

To compare docking results obtained from Autodock Vina and Schrodinger, we ranked the position of each ligand. Each small molecule has 12 ranks, the six members of the ensemble from Autodock Vina and the six members from Glide; we discarded the two highest values, which are the two worst predicted binding affinities, to account for poor protein-conformation/ligand match and then averaged the rank to obtain a ranked list.

For the final docking, we considered the 20 best-predicted binders for each of the conformations (a total of 191 small molecules). To this list, we added ligands with the best-ranked average, further expanding the list to 250 compounds. These compounds were prepared with LigPrep for a final list of 932 different small molecules with different conformations and configurations. Using Glide, we redocked the ligands to each protein conformation with the extra precision (XP) level of theory.

Interaction diagrams were created in Maestro, while protein structure figures were created in Blender 3.0.

4.2.4 Results and discussion

4.2.4.1 Three-dimensional models of human SMUG1

Using the human SMUG1 sequence, we generated a homology model using SWISS-MODEL (SM) web server. The algorithm used the *Xenopus* SMUG1 protein (PDB 1OE4) as the template since the frog protein shares a 65.31% sequence identity with the human protein. The initial model is a homodimer with a high-quality estimation QMEANDisCo of 0.86 ± 0.05 (values closer to one are expected for good models). This initial hSMUG1 model is an asymmetric homodimer.

We refined the structure using another online tool called DeepRefiner (DR). This web server uses deep neural networks to calculate the residue-level errors and subsequently minimizes the errors using an energy-minimization-based restrained relaxation. We optimized the protein as a dimer and also as an individual chain, leading to three new models: chain A from the dimer, chain B from the dimer, and the refined structure from the monomer.

Lastly, the structure obtained by AlphaFold (AF) was also used; this structure has a high confidence measure (average pLDDT 92.39) except on the N-terminus of the protein, where there is a disordered section of 25 residues. The confidence measure pLDDT increases to 97.20 if we ignore the disordered N-termini.

In total, we have six different individual hSMUG1 conformations. Figure 33 shows the difference in backbone RMSD; the conformations obtained by SM are similar to each other, as are the DR conformations. In contrast, the AF model is dissimilar to all the other conformations. Figure 33B shows the pocket conformational diversity, especially observed in the side chain conformations of M84, R243, and the backbone conformation of the loops around the binding pocket (residues A171-T178 and L237-N244). The conformational diversity observed in the pocket is introduced into the docking algorithm when we use all the models.

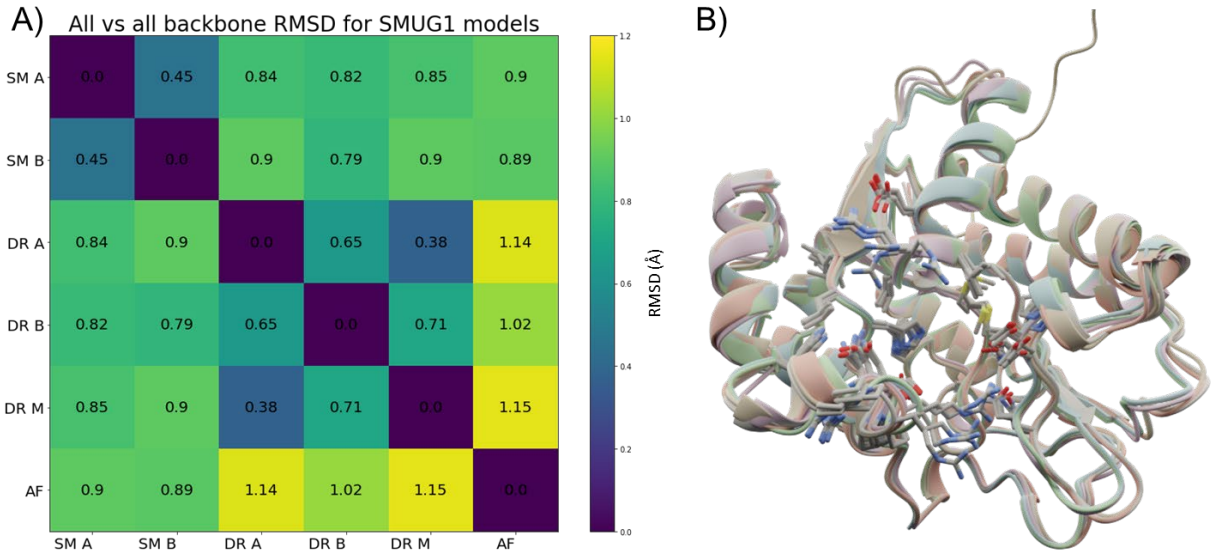


Figure 33. Comparison of SMUG1 conformations. A) All vs. all backbone RMSD of the six different homology models. B) Superposition of the homology models. Showing residues in the binding pocket. The six conformations are SWISS-MODEL chain A (SM A), SWISS-MODEL chain B (SM B), DeepRefiner chain A (RF A), DeepRefiner chain B (RF B), DeepRefiner monomer (RF M), AlphaFold (AF).

4.2.4.2 Small molecule parameterization

To identify novel SMUG1 ligands, we considered the NCI diversity set III (Div3) and diversity set VI (Div6) from the NCI compound sets.[167] These databases were chosen because their compounds are readily available and were tailored to maximize pharmacophore diversity, but they have considerable overlap with most of the compounds being shared between both libraries. We also considered the Enamine Discovery Diversity Set DDS-10; we chose this dataset because the compounds are purchasable from the vendor, and the size of the library (10,240 compounds) was amenable for docking with our computational resources. We calculated all versus all Tanimoto scores to sample database diversity (Figure 34). According to Tanimoto distributions, we can see that the Div3 and Div6 are more diverse compared to the Enamine database, but the Enamine database is about 6.5 times larger.

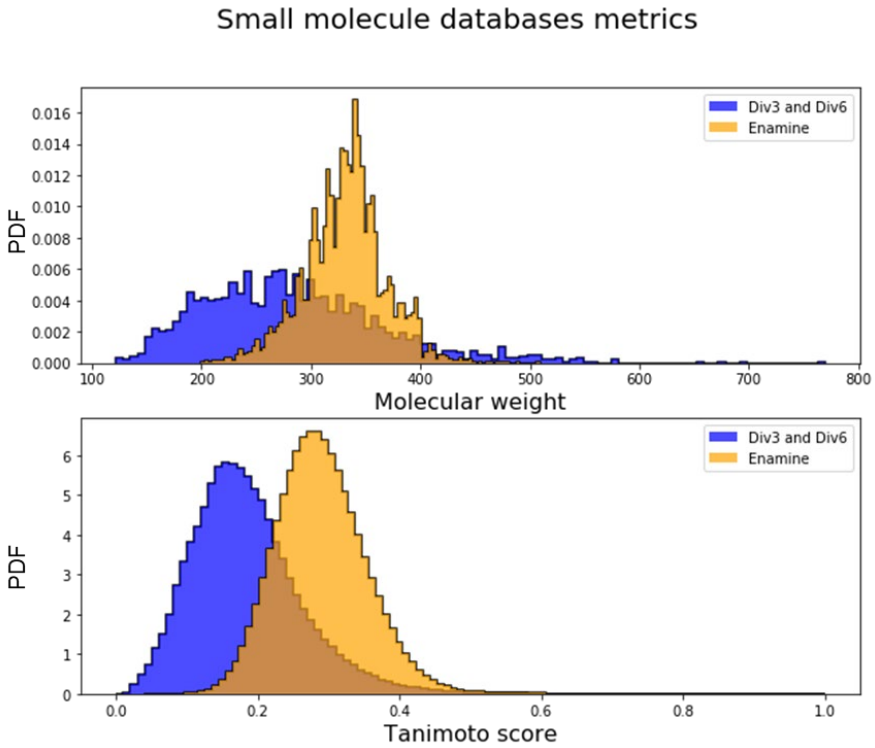


Figure 34. A) Molecular-weight distribution of each database. B) All versus all Tanimoto scores for each database.

To convert the 2D structures from the SDF files to their 3D structures, we used Gypsum[159] with its default parameters. The resulting 3D structures were then processed with MGLTools to generate a final set of 13,421 small molecules ready to use in AutoDock Vina. The original SDF files were also processed with LigPrep, with used the default parameters except that we constrained the maximum number of stereoisomers to 16, this constrain was only applied to the ligands that did not have their configuration defined in the input file. We ended with 39,106 different small molecules for later use in Glide.

4.2.4.3 Ensemble docking of readily available small molecules

Trying to maximize the hit rate for SMUG1, the initial docking of the small-molecule databases was done in two different programs, a technique known as consensus scoring. A significant limitation of docking is the inconsistent performance of different programs. This inconsistency can be addressed by consensus scoring, a method where one combines the results obtained by various docking programs.[168–170] The first tool we used was AutoDock Vina, a free docking tool that may have a steep learning curve for people without computational knowledge. The second was Glide, with standard precision (SP) and default parameters. We used SP for initial docking because of its low computational cost. Glide is a tool that does not have the steep learning curve as AutoDock Vina, but it comes at a high monetary cost.

Visual inspection of the best 20 scored ligands per protein conformation showed that the inspected ligands bound in the active site with no apparent clashes. We created a consensus list to combine the results of our two docking tools. First, for each protein conformation, we assigned a rank to each ligand. Each ligand thus has 12 ranks, six for AutoDock Vina and six for Glide. We averaged the ranks, ignoring the two worst-predicted binders to account for a bad ligand/conformer match. To obtain the list with the best 250 predicted binders, we took the best 20 scoring ligands

for each conformation and each docking program (191 molecules). We added the lowest average rank ligands that were not already on the list to complete it. We took these best-predicted binders and prepared them for follow-up docking, which resulted in 932 different small-molecule models (due to the inclusion of possible stereoisomers). Now using the highest level of theory available in Glide, XP level, we docked the 932 molecules into the six protein conformations.

The compounds' docking scores range from -11.637 to 10.546 kcal/mol, with a mean of -3.730 kcal/mol and a median of -3.931 kcal/mol. I present the distribution of docking scores in Figure 35. Interestingly, the DeepRefiner A conformation has one of the lowest average of docking scores (-3.996 kcal/mol), but none of the poses are on the top-scoring list.

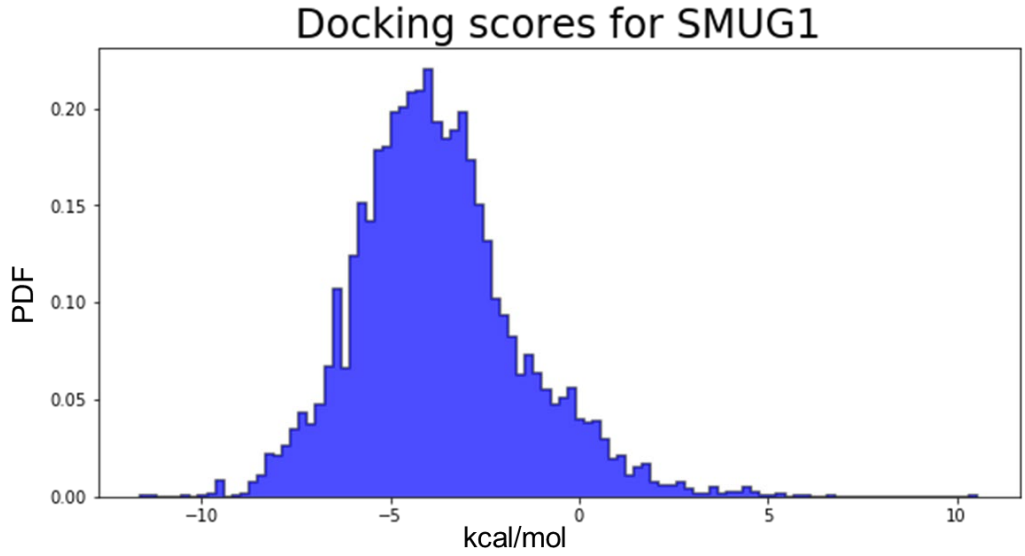
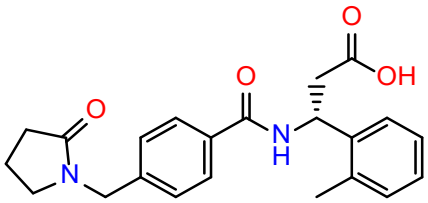
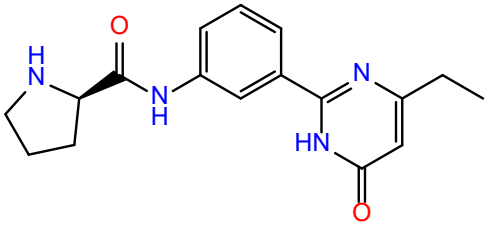
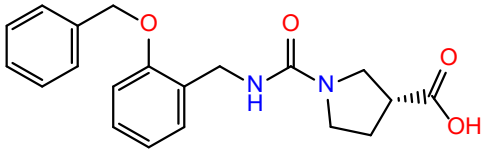
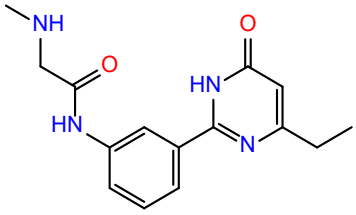
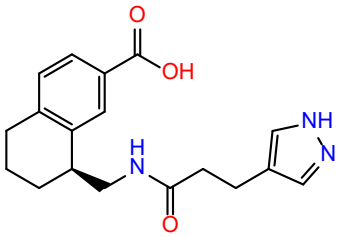


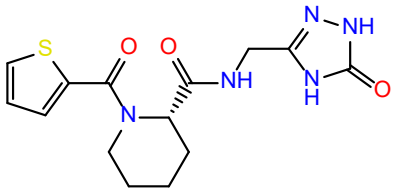
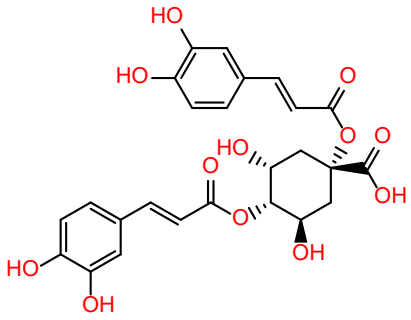
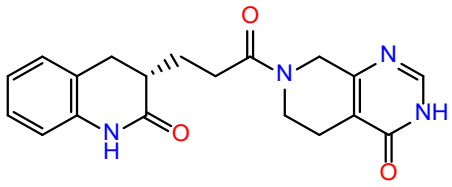
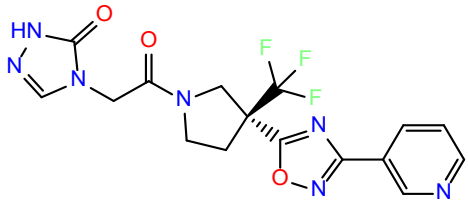
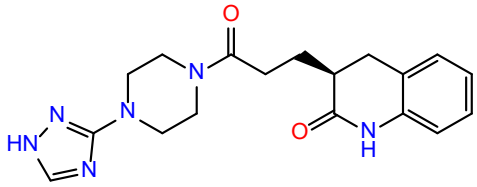
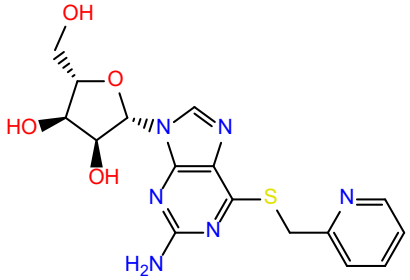
Figure 35. Glide XP docking scores of the 931 best-binder small molecules when docked into the six SMUG1 conformations.

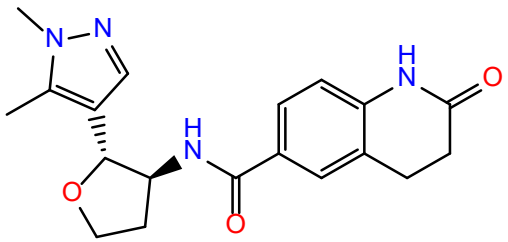
Especially notable was compound NSC 91529, which appeared 16 times with a docking score below -8.5 kcal/mol (the 0.5 percentile cutoff is -8.516 kcal/mol). This compound has four chiral centers and two double bonds. Due to the restrictions we imposed on creating the ligands, only 16 of the 64 possible stereoisomers were generated. Using Maestro, we manually generated

the correct configuration for NSC 91529, a natural product named 1,4-Dicaffeoylquinic acid. We then redocked the correct stereoisomer to the six protein conformations. Table XX shows the 12 top-scoring compounds (99.5 percentile) after pruning NSC 91529's incorrect configuration poses.

Table 3. List of best binders according to Glide docking at the XP level of theory.

Rank (ID)	Compound name	Structure	Protein conformer	Docking score
1	Z1603696798 (Enamine)		DRB	-11.637 (-8.642)
2	Z509792530 (Enamine)		DRM	-10.043
3	Z2108760361 (Enamine)		DRB	-9.862
4	Z2448501887 (Enamine)		DRM	-9.624
5	Z2911212478 (Enamine)		DRB	-9.614

6	Z1332790714 (Enamine)		AF	-8.912
7	NSC 91529 (Div VI)		SMB	-8.793
8	Z1656856947 (Enamine)		AF	-8.774
9	Z2608766234 (Enamine)		SMB	-8.705
10	Z2881982401 (Enamine)		AF	-8.685
11	46385 (Div III)		SMA	-8.671

12	Z1754084691 (Enamine)		DRM	-8.558

4.2.4.4 Assessment of top binders

Computer docking suggests that all the compounds would occlude substrate binding since they all bind in the catalytic pocket. We can observe how ligands completely insert themselves into the pocket of three SMUG1 crystal structures (xSMUG1 PDB 1OE5, and 1OE6[151]; *Geobacter metallireducens* SMUG1 5H9I[171]). This complete insertion was observed in all but three compounds (**3**, **8**, and **12**). To increase the binding affinity of these three compounds, they could be further optimized to include a moiety that interacts with N163, a residue that appears in all the interaction diagrams except these three compounds.

All compounds bind at the entrance of the pocket, pointing towards F89. The region where F89 lies is in charge of discriminating substrate C5 substitutions. In one of the xSMUG1 crystal structures (PDB 1OE6), a 5-hydroxymethyl uracil moiety is positioned at the pocket entrance, providing experimental evidence of the pharmacological potential of the region. The only compound that uses the other side of the pocket entrance is compound **3**; this region could be used to further optimize all other compounds.

The 12 top-ranked compounds participate in some common interactions. In our models, all but one of the compounds form at least three hydrogen bonds with the protein. The residues that participate most in hydrogen bonding are M84 and N85, with six compounds participating in hydrogen bonds with these residues. However, interactions with M84 are more prevalent in the

highest-scoring compounds compared to N85, which suggests that forming an interaction with M84, substantially stabilizes ligands in the pocket. Salt bridges are only present in the first five top-ranked compounds and always form with R124, E135, or R243. These three residues are close to each other, and they line the entrance to the pocket. Finally, five compounds form π-π interactions with the protein, including three that interact with the catalytic H239.

Compound **1** forms six predicted hydrogen bonds with the protein (Figure 36). The compound's carboxyl group is predicted to form a salt bridge with R124, and this group also participates in two hydrogen bonds with N176 and R243. Finally, a hydrogen bond is also predicted to form between compound **1**'s cyclic amide carbonyl group and M84.

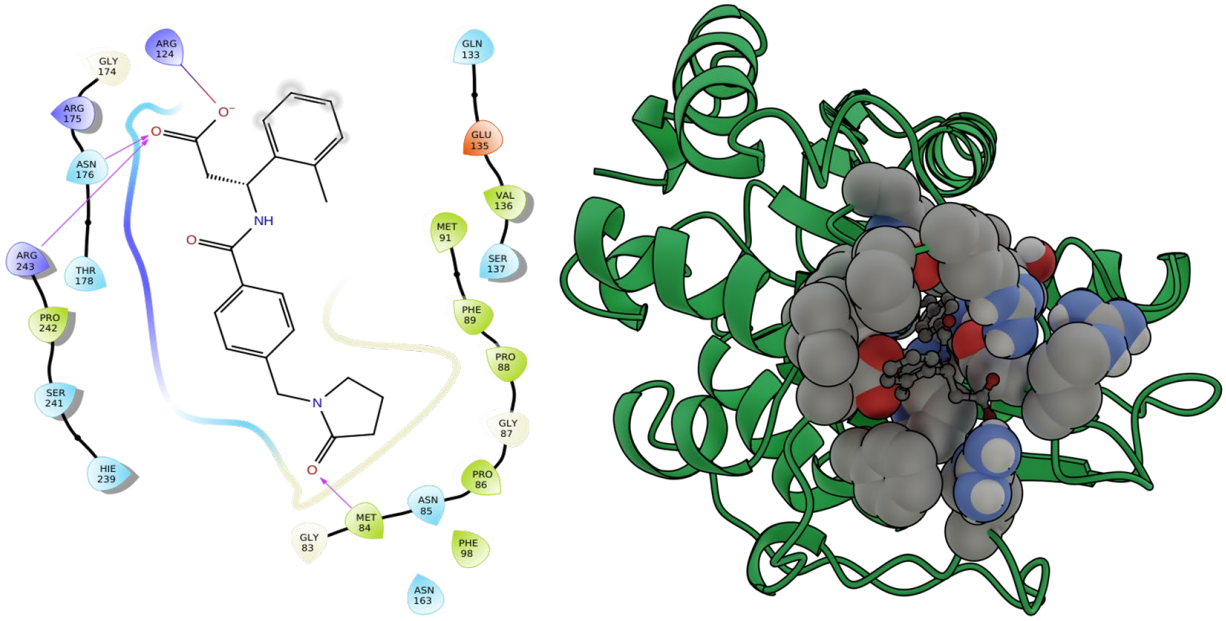


Figure 36. Predicted interactions between hSMUG1 and Z1603696798. A) Interaction diagram; salt bridges are shown as multicolored lines and hydrogen bonds as pink arrows. B) Structure of the protein-ligand complex. Atoms of the pocket residues side chains are shown as spheres, and the ligand is shown as ball and sticks.

4.2.4.5 Conclusions

In the present work, we performed homology modeling of SMUG1 to obtain six different protein conformations. We used three diverse small molecule datasets with compounds available for purchase or delivery. These compounds were docked into six protein conformations using two separate docking programs. We then generated a list of the best 250 compounds, which were docked using Glide at the XP level of theory.

After docking at a higher level of theory, we learned that compounds that can form salt bridges with R124, E135, or R243 tend to have a high predicted binding affinity. Another interaction that seems to increase the binding affinity is the hydrogen bonds with M84 and H239. We recommended 12 compounds for experimental testing because they appear to block the catalytic pocket and/or the entrance to the catalytic pocket.

Below are the contributions each individual made to this project.

- Badiallo Diani - Generated five protein conformers using online tools. Parametrized NCI's diversity set III and VI using MGLtools. Docked two small molecule datasets (NCI Div 3 and NCI Div 6) into five of the protein confirmations (all except AlphaFold, which was obtained after the student left) using AutoDock Vina.
- Ann Wang - Docked a new small-molecule dataset, the Enamine data set, to the six protein conformations using Autodock Vina. Visual inspection of docking results, only the top 20 per conformer. Created a best probable binder list and visually inspected it. Created a consensus list for the overall best binders.

- Erich Hellemann- Mentored undergraduate students. Using Glide, docked the three small-molecule data sets to the six protein confirmations. Performed the last docking step with the best 250 compounds into the six protein confirmations using Glide at the XP level of theory.

4.3 Small molecule binding to TEM-1 β-lactamase's cryptic pocket changes side chain dynamics, leading to inhibition

4.3.1 Justification

This project started as an undergraduate research experience for Amrita Nallathambi. The student had plenty of experience with computational tools and also had experience with computational biophysics methods. The project aimed to determine the atomistic mechanism by which a small allosteric molecule inhibits TEM-1 β-lactamase activity to confer resistance to lactam-containing antibiotics. The student began with a literature search to understand the protein's biology and relevance and learn the basics of computational biophysics. This reading was part of the Durrant laboratory's literature club, which I initiated and guided. For the project, the student had to parametrize the protein and the small molecule for molecular dynamics simulations. Once the parameterization was done, she ran the molecular dynamics simulations at the Center for Research and Computing at the University of Pittsburgh. She then analyzed the simulations by calculating the RMSD of the protein and ligand and the RMSF of the protein. She also performed cluster analysis, PCA, side chain sampling, and network analysis. Regrettably, when starting to write the manuscript, we encountered an issue with the parametrization of the protein and so had to repeat the simulations and analysis. I did all the work presented here and wrote the final manuscript for publication.

4.3.2 Introduction

The discovery of antibiotics in the early 20th century was a significant driver in decreasing morbidity and mortality in the human race. However, decades of indiscriminate prophylactic and therapeutic use in agriculture, human health, and research have put incredible evolutionary pressure on microorganisms forcing them to adapt and evolve to resist these compounds.[172–174] Antimicrobial resistance (AMR) is recognized by the World Health Organization (WHO) as one of humanity's top ten global public health threats.[175] β-lactam antibiotics, which are the most prescribed and sold class of antibiotics,[176] were discovered in the early 20th century by Alexander Fleming. These compounds block the last step of bacterial cell wall formation by targeting DD-transpeptidases, also known as penicillin-binding proteins (PBPs). These proteins are responsible for the cross-linking of peptides in peptidoglycan synthesis. β-lactam antibiotics, as their name suggests, have a lactam ring that bacteria have evolved to open by hydrolysis. Once the lactam ring is open, the antibiotic is rendered inert, and its antimicrobial activity is blocked.

Proteins that open the lactam ring are named lactamases and are divided into four classes (A through D). Class A, B, and C are serine hydrolases, and class D are metallo-lactamases. TEM-1, a class A lactamase member, was first identified in 1963 when it was isolated from penicillin-resistant bacteria.[177] This discovery led to the development of new lactam antibiotics, which inevitably led to the appearance of new lactamases. This cycle of new drugs, followed by the evolution of resistance, is known as the "β-lactamase cycle." [178,179] Currently, there are over 240 TEM variants, as reported in the Beta-Lactamase DataBase.[180]

Structural studies of class A β-lactamases have shown that they have two tightly packed domains, a primarily α helical domain, and an α/β-domain. Three α-helices and five anti-parallel β-sheets forms the α/β-domain. The mostly α helical domain contains nine helices, most

catalytically relevant residues, and most of the catalytic pocket lining residues. The acylation and deacylation reactions required to open the antibiotic lactam ring happen through an activated serine. Activation of the catalytic S70 is proposed to go through the general base E166. This interaction is mediated by a conserved water molecule since E166 and S70 are too distant. This water molecule is observed in many class A lactamase crystal structures.[179,181]

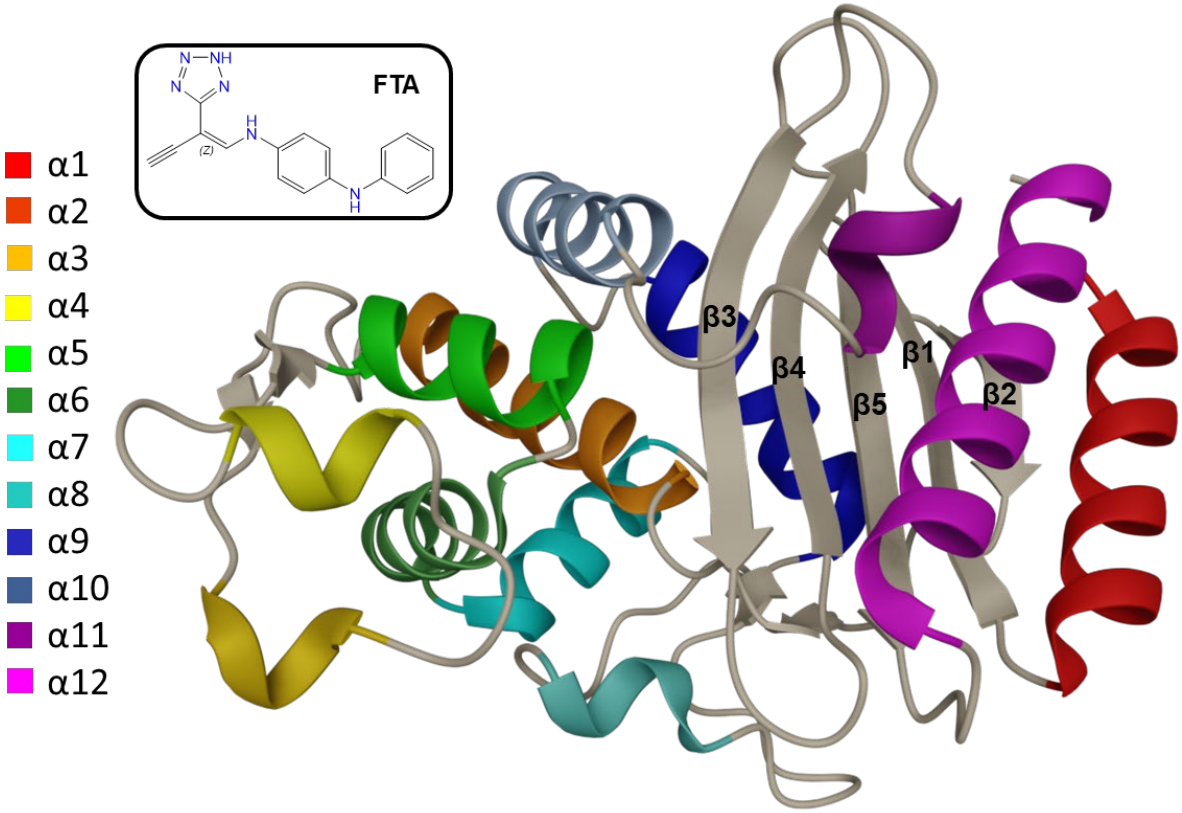


Figure 37. TEM-1 β-lactamase with bound allosteric inhibitor FTA. Structure from PDB 1PZP.

An allosteric pocket in the α/β-domain was discovered by Horn and Soichet[182] when they found TEM-1 inhibitors that did not block the orthosteric pocket. This cryptic pocket is formed when α11 rotates and moves away from α12. Allosteric inhibition of β-lactamases is an attractive avenue to tackle the problems arising from the "β-lactamase cycle" because an allosteric inhibitor can be highly selective, and in this case, target a highly conserved region of the protein.

In the present work, I use molecular dynamics (MD) simulations of TEM-1 with the allosteric inhibitor 3-(4-phenylamino-phenylamino)-2-(1H-tetrazol-5-yl)-acrylonitrile (FTA) bound (*holo*) and without the bound inhibitor (*apo*), for a total of ≈3 μs of simulation time. MD simulations give us an atomistic view of the changes in TEM-1 dynamics that FTA binding induces. MD, coupled with network analysis, is an efficient way of elucidating allosteric pathways.[36,37] I applied these principles to discover that FTA may elicit population changes in R244, which would weaken the binding of a β-lactam ring containing antibiotics. The binding of the inhibitor also changes inter-protein communications, which could further help with inhibition. Our simulations also revealed the opening of a larger pocket and the complete burial of the small allosteric molecule. This pocket is similar to the one reported for N,N-bis(4-chlorobenzyl)-1H-1,2,3,4-tetrazole-5-amine (PDB 1PZO[182]). I suggest that medicinal chemists could exploit the pocket observed in the alternative binding mode to develop novel TEM inhibitors.

4.3.3 Methods

4.3.3.1 File preparation

Crystal structures of the *apo* (PDB 1ZG4[183]) and *holo* (PDB 1PZP[182]) complexes of TEM-1 were downloaded from the Protein Data Bank. The complex structure includes two copies of the small molecule 3-(4-phenylamino-phenylamino)-2-(1H-tetrazole-5-yl)-acrylonitrile (FTA), one of which is bound in the cryptic pocket being investigated. The second FTA lies at a protein-protein interface, the product of crystallographic contacts, and so was removed. There is a difference of three residues in the sequences of both proteins. Using ChimeraX,[184] I mutated residues I84V, N100R, and V184A of the *holo* protein, so the only difference would be the ligand and initial coordinates. The protein structures in the PDB files were then checked using MolProbity[185] and

corrected for side chain flips. After visual inspection, only residues Q39 and N276 for the *holo* proteins were deemed necessary. A general visual inspection of crystallographic structures and electron density maps was also performed. Using the PDB2PQR[53] web server, which employs the PROPKA[54] algorithm, I added hydrogens to all the proteins at pH 7.

The ligand was individually hydrogenated using the reduce function from Leap, a program from the AmberTools20 package.[55] Partial charges were obtained with the default parameters at the AM1-BCC level of theory.

The smallest water box that completely enveloped each protein was optimized by rotating the protein or protein-ligand complex. The proteins were then solvated in this position with a water box that extended 10 Å in every direction beyond the protein. The systems were neutralized using Na+ counter ions, and Na+ and Cl- ions were used to achieve a 0.15 M solution. The systems were parametrized in Antechamber using Amber's ff14sb,[56] GAFF,[186] and TIP3P[57] force fields for protein, small molecule, and water, respectively.

4.3.3.2 Molecular Dynamics simulations

Molecular dynamics (MD) simulations were performed with the NAMD 2.13 package.[58] The systems were minimized in a four-step process of 5000 steps each. First, all hydrogens were relaxed; then hydrogen atoms and water molecules; continued by hydrogen atoms, water molecules, and protein side chains; to end with all atoms being relaxed.

Equilibration was performed in the NPT (isothermal–isobaric) ensemble at 310K, using long-range Particle Mesh Ewald electrostatics (14 Å cutoff), the Nosé-Hoover method to keep a pressure of 1.01325 bar, and Langevin dynamics (damping constant 5 /ps). For the *apo* simulations, a four-step equilibration was performed. In each 0.25 ns step, I applied harmonic constraints to the protein backbone atoms, which were gradually relaxed in each step (1.0, 0.75,

0.5, 0.25, 0.00 kcal/mol/Å^2, respectively). For the *holo* simulations, a single 1 ns equilibration step was performed with a one fs time step. All simulations used the SHAKE[127] algorithm.

Once equilibration finished, three different production runs per system were set up with identical conditions to start from the last frame of equilibration. Three different production runs of lengths 250ns, 250ns, and 500ns were obtained for each system. Using MDAnalysis, the backbone RMSDs of each trajectory were calculated. This measure suggested that during the first 10ns, some of the systems continued to equilibrate, leading me to remove this portion of the simulation before carrying out further analysis. For consistency, each run was similarly trimmed, and these truncated runs were used in the analyses described in the rest of the paper. Finally, the simulations were aligned by their backbone atoms using MDAnalysis (V 1.0.0).[60,61]

4.3.3.3 Induced Fit Docking, MM-GBSA, and binding pose metadynamics

I clustered the *holo* simulation in which the ligand inserted itself more into the protein. I performed two clustering analyses. The first centered on the part of the simulation with the ligand in the crystallographic pose and the second on the "horizontal" pose (see further description below).

I assembled a conformational ensemble comprising the crystallographic structure, the centroid of the most populated cluster for both clustering analyses, and a single frame from the simulation with the ligand in its "horizontal" pose. I prepared all the conformations with Schrödinger's protein preparation wizard.[164] To prepare FTA for docking, I used LigPrep[163] from Schrödinger with the OPLS4 forcefield.[187] I then used the induced fit docking[188,189] module in Schrödinger to re-dock FTA to all protein conformations with extended sampling. This docking procedure considers the flexibility of the protein and the ligand, something regular docking does not. The poses with the lowest binding energy for each protein conformation (crystallographic and

"horizontal") were further processed with Prime-MMGBSA. The lowest binding energy poses from IFD and MMGBSA, the crystallographic pose, and the extracted "horizontal" pose were then used in the binding pose metadynamics module from Schrödinger.[190,191]

4.3.3.4 Analysis of molecular dynamics simulations

Visualization of the systems and simulations was done with VMD.[192] MDAnalysis[60,61] or CPPTRAJ[63] were used to analyze the trajectories and calculate RMSD, RMSF, and distances. Side chain angles for some residues of interest were calculated and plotted using the Janin class of MDAnalysis. Clustering of Janin plots was performed with K-Means as implemented in Scikit-Learn.[193]

The secondary structure of TEM-1 was predicted using the online server Stride[194] with PDB 1ZG4 as the template.

4.3.3.5 Network analysis

Further analysis of the relationships between amino acids within the protein was conducted using the MD-TASK suite of modules.[128] Dynamic cross-correlation (DCC) between the residues for each system was calculated using every tenth frame. The results of the DCC analysis yield a matrix that describes the degree of correlation between Cα atoms, and each element of this matrix is calculated as:

$$C_{ij} = \frac{\langle \Delta r_i \cdot \Delta r_j \rangle}{\sqrt{\langle \Delta r_i^2 \rangle} \cdot \sqrt{\langle \Delta r_j^2 \rangle}}$$

I performed analysis on the concatenated and aligned individual simulations. This analysis gives a value between -1 and 1 for each pair of residues. Values close to one are residues whose movement is correlated, while values close to -1 are for residues moving in an anticorrelated way.

Additionally, I calculated the betweenness centrality of each residue (BC), and the average shortest path to each residue (L). For these analyses, the protein is represented as a set of nodes that are said to interact if they are near each other. An edge connects Cβ (Cα for glycine) carbons within a cutoff of 7.0 Å.

The BC of a node denotes the number of shortest paths that pass through that node when calculating all vertices' shortest paths to all other vertices. BC measures how important a given node is for communication within the network (protein). L is the sum of all the shortest paths to the residue divided by the number of residues minus one. ⟨L⟩ describes the accessibility of the residue within the protein network.

The averages and changes in BC and L across the trajectories were also determined and used to identify residues with distinctly different behavior in the *apo* and *holo* simulations.

4.3.4 Results and analysis

To understand the mechanism by which the small molecule FTA inhibits the catalytic activity of TEM-1, I performed molecular dynamics (MD) simulations of three different systems. The first system is wild-type (WT) TEM-1 in its *apo* form. The initial coordinates of the *apo* protein were obtained from PDB 1ZG4. The second system is the *holo* mutant system, which I downloaded from PDB 1PZP. This structure, reported by James R. Horn and Brian K. Shoichet,[182] shows the binding between TEM-1 and FTA a small molecule allosteric inhibitor. The crystallographic complex contains two copies of the small molecule inhibitor, one in a cryptic

pocket that opens up when α11 moves away from α12. The second lies at a protein-protein interface, a product of crystallographic packing. Visual inspection of the electron density and crystal packing also showed contacts between the tetrazole ring in FTA and one of the adjacent proteins in the crystal structure. I manually deleted the FTA molecule at the protein-protein interface and called this the *holo* mutant complex because the protein differs by three amino acids compared to the WT system. In WT, we have I84, N100 and V184, but in the mutant protein, these are valine, arginine, and alanine, respectively. To generate the third system, the *holo* WT system, I used ChimeraX to computationally mutate these residues in the *holo* mutant complex to the residues in the WT protein.

4.3.5 Ligand binding increases protein stability

The systems were prepared, minimized, and equilibrated as described in the methods section. The equilibrated systems were run in the NPT ensemble, and each system had three replicates totaling one μs of simulation time per system. The simulations were aligned, and I calculated the RMSD of backbone atoms for each system, showing the protein to be rigid. This rigidity agrees with Nuclear Magnetic Resonance (NMR)[195] and previous Molecular Dynamics (MD) studies[196–198] which have shown a highly rigid protein as observed by a high order parameter for the whole protein ($\langle S^2 \rangle = 0.90 \pm 0.02$) and low RMSD values in the simulations.

The *apo* simulation showed a higher degree of mobility ($\langle \text{bbRMSD}_{apo} \rangle = 1.15 \pm 0.17$ Å) according to its backbone RMSD compared to the *holo* simulations ($\langle \text{bbRMSD}_{holo\ \text{wt}} \rangle = 0.91 \pm 0.10$ Å and ($\langle \text{bbRMSD}_{holo\ \text{mut}} \rangle = 0.94 \pm 0.09$ Å).

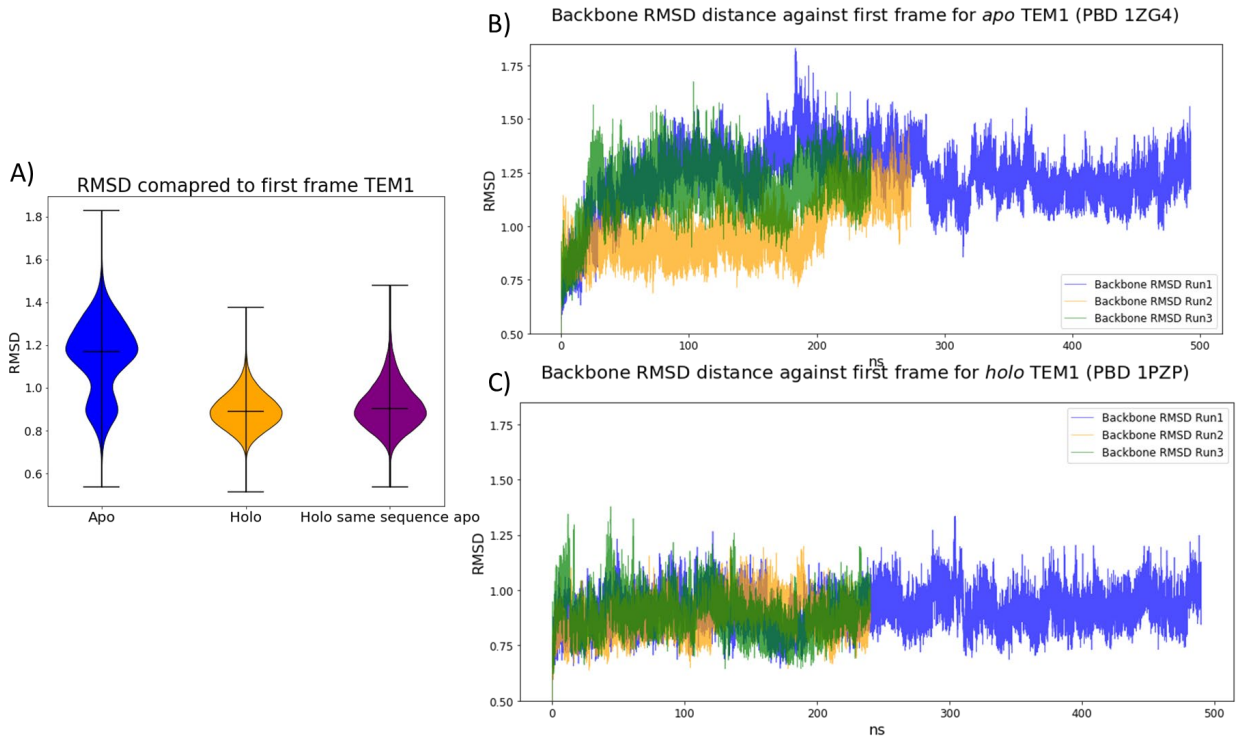


Figure 38. Comparison of protein confirmational sampling between *apo* and *holo* simulations. A) Violin plots of backbone RMSD compared to the first frame, wild-type *apo* and *holo* simulations. B) RMSD to the first frame for the *apo* TEM1 simulations. C) RMSD to the first frame for the *holo* TEM1 WT simulations.

To identify the protein regions with greater flexibility, I calculated RMSF values per residue for all three systems using each residue's center of geometry (CoG). I calculated ΔRMSF values per residue (*apo* - *holo*) and observed overall higher flexibility in the *apo* simulations (Figure 38). Our results are similar to those obtained for the simulations of TEM-1 bound to BLIP, a known protein that inhibits TEM-1, reported by Meneksedag and coworkers. They see a reduction of mean square fluctuations in the inhibitor-bound systems.[199,200] In our case, I see a reduction from $\langle RMSF \rangle_{apo} = 0.78 \pm 0.39$ Å for the *apo* system to $\langle RMSF \rangle_{holo\ WT} = 0.71 \pm 0.32$ Å and $\langle RMSF \rangle_{holo\ mut} = 0.70 \pm 0.31$ Å for the *holo* WT and mutant, respectively. In contrast, the results differ from those of Shozeb Haider and coworkers.[201] These authors found that the FTA ligand binding decreases protein stability, as observed by increased RMSD and RMSF, especially

in three regions. The difference between their simulations and ours could explain the discrepancy between the results. The authors of this publication obtained their *apo* protein by deleting FTA from the complex and equilibrating the system. We ran the simulations in the isobaric-isothermal ensemble (NPT) while they ran in the canonical ensemble (NVT). Finally, they used a different MD engine, ACEMD, while we used NAMD.

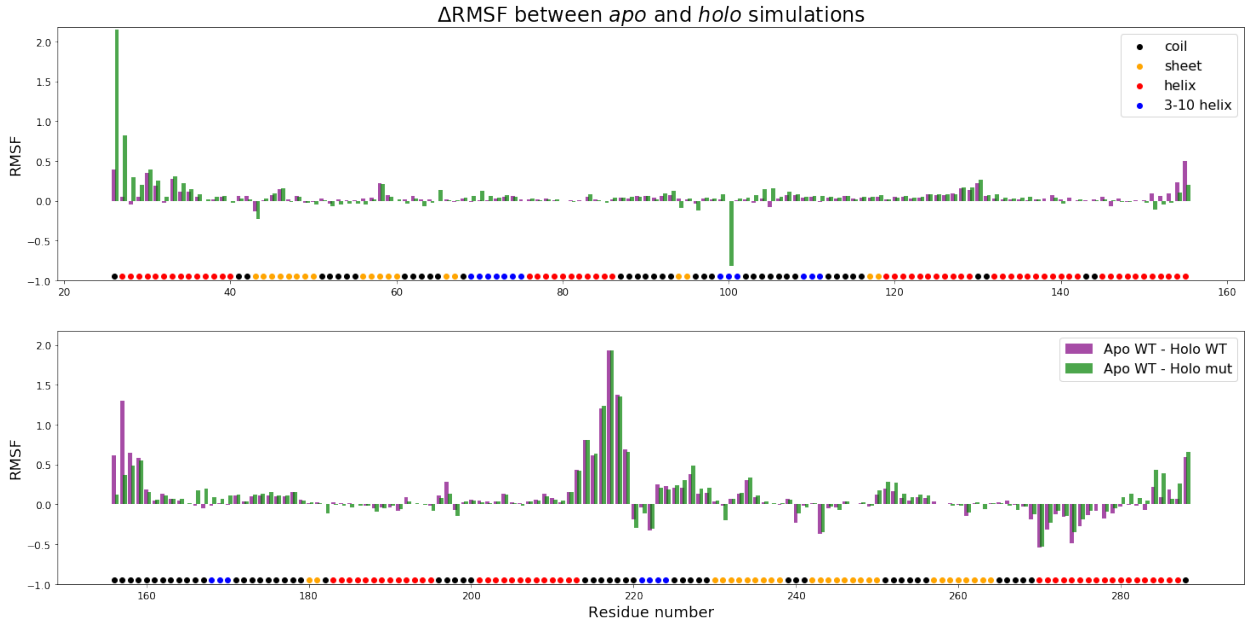


Figure 39. Difference in per-residue RMSF (center of geometry) between *apo* and *holo* simulations (*apo* – *holo*).

There are two regions where the *apo* simulation is substantially more flexible than the *holo* simulations. The first region is one of the hinges connecting both the primarily α helical domain, and an α/β-domain. This hinge is comprised of the loop connecting helices α10 and α11, with α11 having direct contact with FTA. The second region of higher flexibility for the *apo* simulations is comprised of residues after α7 (residues 155-162). α12 has a slightly higher RMSF in the *holo* sims, especially towards the N-terminus of the helix (residues 269-279). This helix is also in contact with FTA, but contacts with the crystallographic pose are on the C-terminus of the helix. Another region with higher fluctuations in the *holo* systems is three residues in α11 (residues 220-

222), with L221 lining the binding pocket. Differences between *holo* simulations are small (less than 0.3 Å) except for the coil after α7, where the mutant is more flexible. There is one RMSF outlier, residue 100 for *holo* mutant; this residue is one of the mutations observed in 1PZP (N100R), but if we compare RMSF of only Cα atoms, we see a 0.026 Å difference between both *holo* simulations. Catalytic residue S70 and the general base E166 have lower RMSF values for the *holo* mutant simulations, but the changes fall below 0.3 Å. In general, the effect of FTA seems to make the protein more rigid, making it less likely to adopt different conformations.

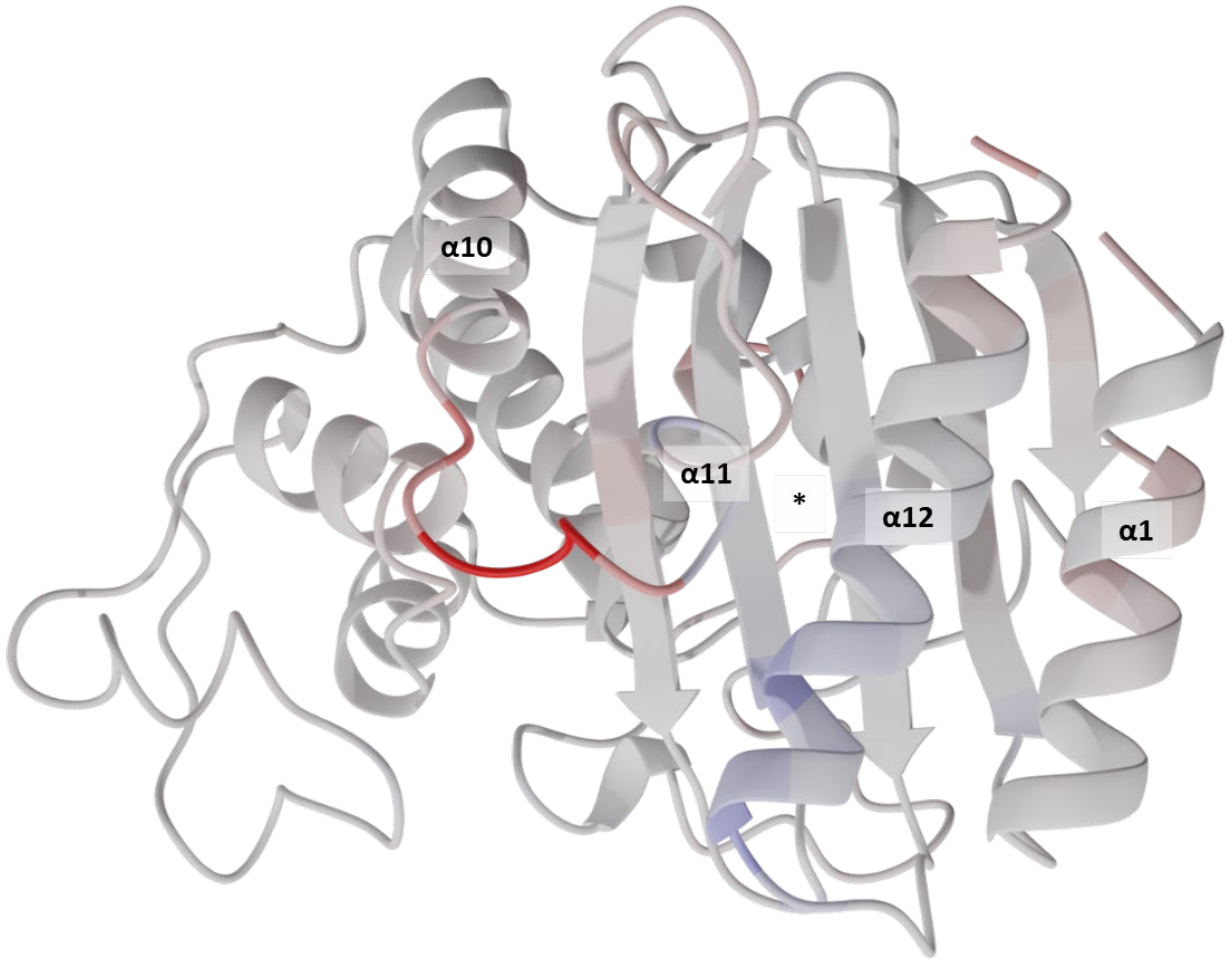


Figure 40. RMSF difference between *apo* and *holo* WT simulations. Red residues have higher RMSF in the *apo* simulations, and blue residues have higher RMSF in the *holo* WT simulations. RMSF differences are

projected onto the first frame of the *apo* simulation. The location where the ligand binds in PDB 1PZP is marked with an asterisk.

To observe differences in essential dynamics between the three systems, I performed Principal Component Analysis (PCA) on the backbone atoms of the simulations. This analysis shows that the first two components account for about 33.8% of the data variance and that the *holo* simulations have only one distinct well compared to the two wells in the *apo* simulations (Figure 41). If we go to the third and fourth components, for an accounted total variance of 48.8%, we also observe more than one well for the *apo* simulation. These differences in distributions agree with the RMSD and RMSF results in that we generally observe reduced conformational sampling in the *holo* simulations.

In the *holo* simulations, the first component captures the allosteric pocket's opening, with α11 moving away from α12. On the other hand, in the *apo* simulations, this component describes the movement of the loop connecting α11 and α12, and the loop after α7. These are the same regions of increased flexibility in the *apo* simulations.

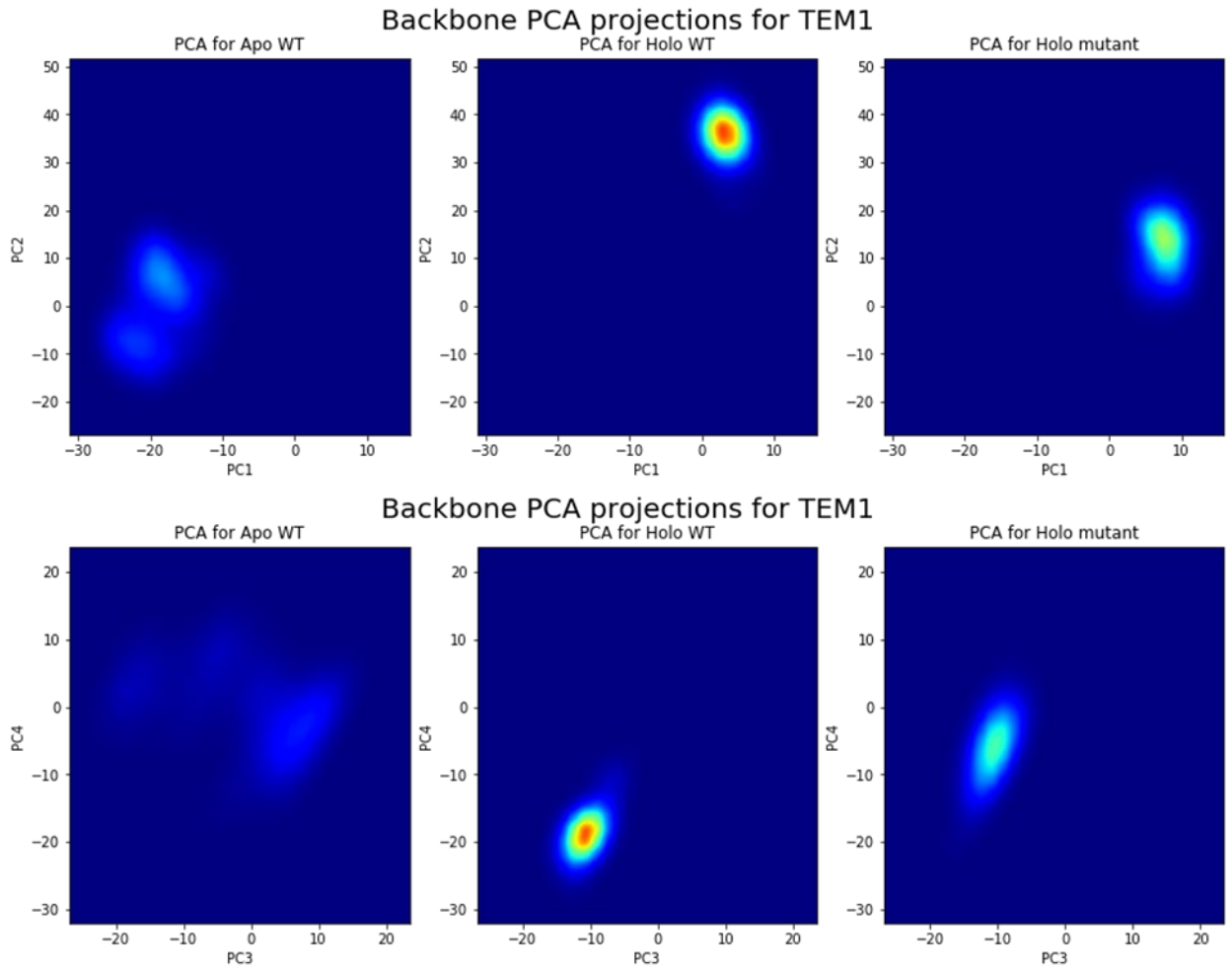


Figure 41. PCA projections for the TEM-1 simulations. Top row shows the 1st and 2nd principal components. Bottom row shows the 3rd and 4th components.

4.3.6 *Holo* WT simulations suggest a novel alternate FTA binding pose

I observed that in one of the *holo* WT simulations, the ligand buried itself completely between α11 and α12, as shown in Figure 42. For that ligand insertion to happen, the secondary amine connecting both aromatic rings has to invert. The amine's hydrogen pointing towards the N-terminus of α12 breaks its interaction with I279's carbonyl and inverts itself to point towards the C-terminus of α12. The L221 side chain then packs closer to the β-sheet, further opening the

pocket, and α12 rotates counterclockwise. This pocket opening allows the tetrazole moiety to insert itself between α11 and α12. From now on, the inserted pose is referred to as the "horizontal" pose.

To understand the two different binding poses, I performed packing analysis using 1PZP's electron density map. The TEM-1 crystal structures obtained by Horn and coworkers were obtained by crystal soaking, a technique shown to present in some cases different binding poses than the slower, more difficult, and more accurate cocrystallization method.[202] This effect is observed more dramatically when there are crystal contacts with the ligand and the neighboring unit, as observed in trypsin bound to an aminopyridine derivative (PDBs 6QL0 and 6T5W). In our case, FTA forms a hydrogen bond between one of the nitrogens in the tetrazole moiety and K192 (distance 3.067 Å) of the neighboring protein in the crystal structure. This significant interaction, a product of crystallographic packing, could stabilize the crystallographic pose. A structure of SHV-1 β-lactamase (68% sequence identity) with a ligand in the same allosteric pocket (PDB 4ZAM),[203] shows a similar pose. This ligand also interacts extensively with its neighbor in the crystal structure, in this case, four hydrogen bonds.

Figure 42. Alternate binding pose ("horizontal") identified in one of the *holo* WT production runs.

The binding observed in the "horizontal" pose is similar to the one observed in another crystal structure obtained by Horn and Shoichet (PDB 1PZO). Instead of FTA, this structure has two copies of N,N-bis(4-chlorobenzyl)-1H-1,2,3,4-tetrazole-5-amine. The location of the tetrazole ring in the "horizontal" pose is close to where one of the tetrazole rings lies in the 1PZO structure, giving some experimental validity to the "horizontal" pose. The opening of the pocket that could accommodate the ligand in the "horizontal" position was also previously predicted by elastic network models.[204]

4.3.7 Ligand pose assessment

To assess which ligand pose is most likely, I performed induced-fit docking (IFD). I used the crystallographic structure, the centroid of the most populated cluster using only the part of the

simulation that contains the crystallographic pose, a single frame with the ligand in its "horizontal" pose, and the centroid of the most populated cluster using only the part of the simulation that contains the "horizontal" pose. Results of IFD can be seen in Table 4. At least 15 poses of each system were used to calculate binding affinities with Prime MM-GBSA. The poses with the best scores for IFD and MM-GBSA were then used for Binding Pose Metadynamics studies, along with crystallographic pose and the single frame from MD simulations with the "horizontal" pose. The crystallographic system was the only pose with both the highest IFDScore and MM-GBSA predicted binding affinities. BPM results are also presented in Table 1; these results show higher stability for the "horizontal" pose with an average score of 2.113 ± 0.891 Å, compared to the crystallographic pose average of 2.975 ± 0.151 Å. By the Shapiro-Wilk test for normality, we do not have normally distributed data for BPM scores (0.223 and 0.07 for crystallographic and "horizontal" poses), so I performed a Mann-Whitney U rank test. The Mann-Whitney U test did not allow us to discard the hypothesis that both samples come from a different distribution ($p = 0.09$). We could have obtained these results because we are underpowered (we only have 4 and 5 samples). However, the observed effect size for the stability of the ligand due to the change in pose is large ($\eta^2 = 0.24$). An experiment that could be performed to assess the validity of the proposed pose would be chemical shift perturbation in solution-state NMR; this way, no crystal contacts would be observed, avoiding the stabilization of the tetrazole ring outside the protein pocket.

Table 4. Ligand-pose assessment using induced-fit docking (IFD), MM-GBSA, and binding-pose metadynamics (BPMD).

	IFDScore (Docking Score)	Best MM-GBSA (best IFDScore)	BPM score (not from IFD)
Crystallographic	-11653.68 (-9.398)	-63.56 (same structure)	2.723 (3.039)
Centroid	-11561.41 (-9.311)	-64.00 (-59.24)	3.017, 3.123

crystallographic			
“horizontal”	-11554.00 (-9.437)	-81.35 (-74.94)	1.264, 1.917 (1.646)
Centroid “horizontal”	-11524.65 (-7.542)	-53.75 (-61.72)	1.905, 3.831

4.3.8 Residue side chain population differences affect ligand binding and stabilization

The mechanism by which FTA inhibits TEM-1 is allosteric. The center of geometry of the ligand is located 21 Å away from the Cα of the catalytic S70, and the closest any FTA atom comes to the same Cα is 15.46 Å, according to the crystallographic structure. To understand how side chain dynamics affect protein function, I calculated Janin plots (χ1 versus χ2) of some critical residues (K73, Y105, E166, W229, K234, R244, and R275).[182,205,206]

Tyrosine 105, a residue that lines the catalytic pocket and is crucial for ligand recognition and stabilization,[206] undergoes a population shift in our simulations. The most visited state for the *apo* simulations has the phenolic ring pointing towards the solvent, this conformation is observed in ligand-bound structures of TEM-1 and TEM-1 bound to BLIP, a known protein inhibitor.[207] This conformation is sampled 43.65% of the time in our *apo* simulations, but only 34.58% and 21.00% in the *holo* WT and *holo* mutant, respectively. In the *holo* simulations, a conformation with the phenolic ring packet against P107 is more populated than in the *apo* simulations, going from 11.79% for the *apo* to 29.53% for *holo* WT and 42.59% for *holo* mutant. The population changes observed in Y105 side chain conformation due to FTA binding could affect TEM-1’s ability to recognize ligands in the orthosteric pocket.

The residue with the biggest change in the side chain conformation according to the calculated Janin plots is R244. This arginine is close to FTA in the crystallographic structure (~11 Å), and has been proposed to stabilize the binding of lactams in the catalytic pocket.[208,209] The

conformational change this residue presents due to FTA binding was proposed as the underlying mechanism by which FTA inhibits TEM-1.[182] R244 has a different starting conformation in the *apo* and *holo* simulations; in the *apo* system, the side chain points towards the orthosteric pocket. In contrast, in the *holo* conformation, the side chain points towards the C terminus of α12 (Figure 43).

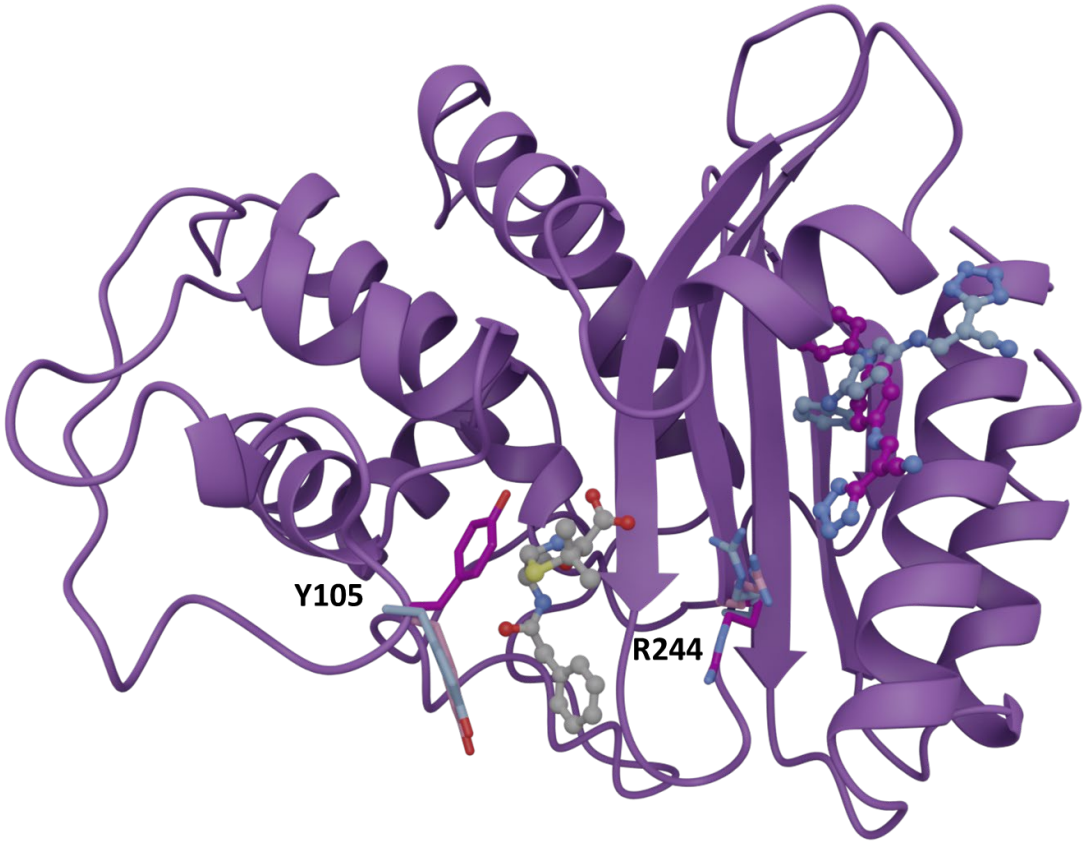


Figure 43 FTA binding influences Y105 and R244 side chain dynamics. Here I show representative conformations of Y105 and R244 side chain conformations, in dark purple, a representative conformation with FTA in the "horizontal" conformation. In light blue, a conformation from the simulations close to the 1PZP crystallographic R244 conformation. The open form of penicillin G in the orthosteric pocket was obtained by superimposing PDB 1FQG to one of the conformations from the *holo* WT simulation.

Arginine 244 also shows a significant shift in the population of side chain conformations over the course of the simulations. The side chain does not deviate much from its starting position

in the *apo* simulations, with 84.32% of the simulation staying in that conformation. The *holo* simulations spend most of their time in a conformation that has the side chain closer to α12. The *holo* WT simulations spend 37.24% of the time in that closer conformation and the *holo* mutant 39.19%, compared to 2.97% in the *apo* simulation.

The R244 side chain conformation observed in the *holo* crystallographic structure is seldom sampled in the *apo* simulations (0.67% of the time), compared to 16.50% and 3.21% for *holo* WT and mutant, respectively. The RMSF observed for the residue has a substantial increase in *holo* systems compared to the *apo* simulations, with a 79.01% and 74.45% increase in RMSF for *holo* WT and mutant, respectively (0.4689 Å for *apo*, 0.8394 Å for *holo* WT and 0.818 Å for *holo* mutant); this increase of fluctuations in the residue is in agreement with the clustering of Janin results.

Interestingly, when I only take the portion of the simulations where the ligand is in the "horizontal" pose, the χ1 and χ2 R244 conformation always match that of the ligand-bound crystal structure.

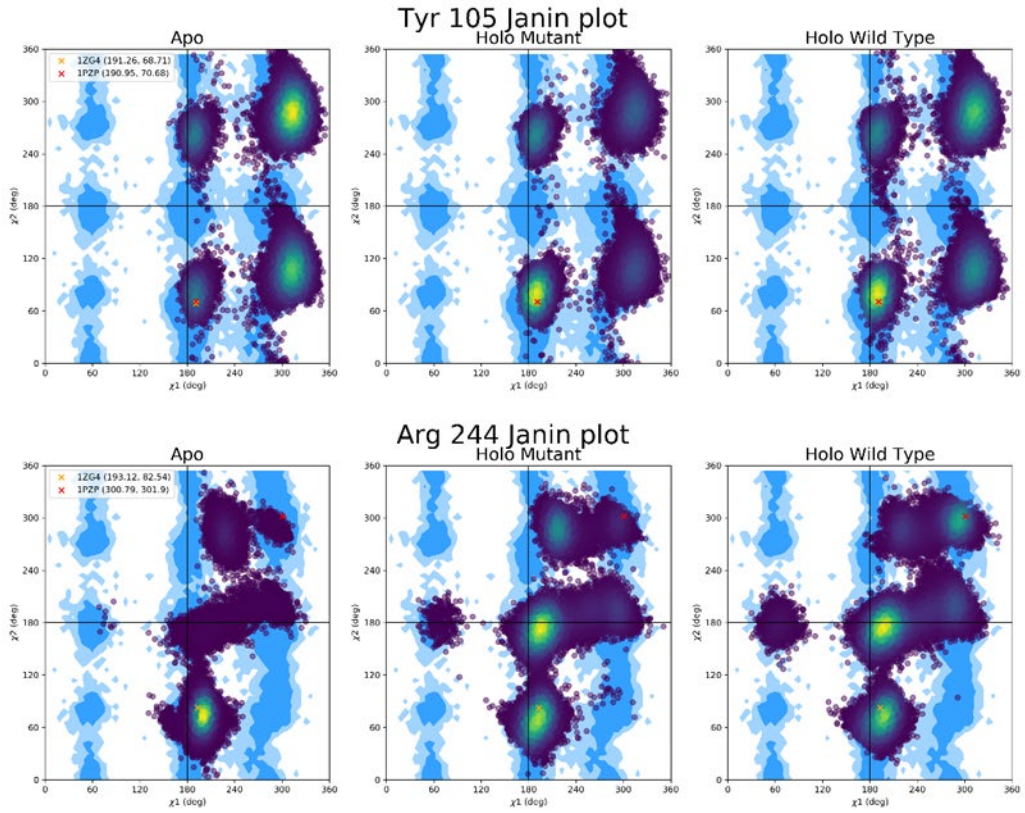


Figure 44. Janin plots for side chains with the most significant population shifts. The upper line contains Janin plots for tyrosine 105, a residue responsible for ligand recognition. The lower line includes Janin plots for arginine 244, a residue responsible for ligand stabilization in the orthosteric pocket.

I calculated the distance between the R244Cζ and the orthosteric pocket (S70Cα). I observed a nearly Gaussian distribution for the *apo* system, but we observed a bimodal distribution for both *holo* systems. I fitted bimodal distributions to all the systems and barely observed the second distribution in the *apo* system. The average distance between both nodes in the distributions is 1.46 Å, going from 10.14 Å to 11.60 Å. As can be seen by all these results, Arg 244 has vastly different dynamics when FTA binds to TEM-1, which makes the residue more flexible and orients its side chain away from the orthosteric pocket, thus precluding its stabilizing effect on antibiotic binding.

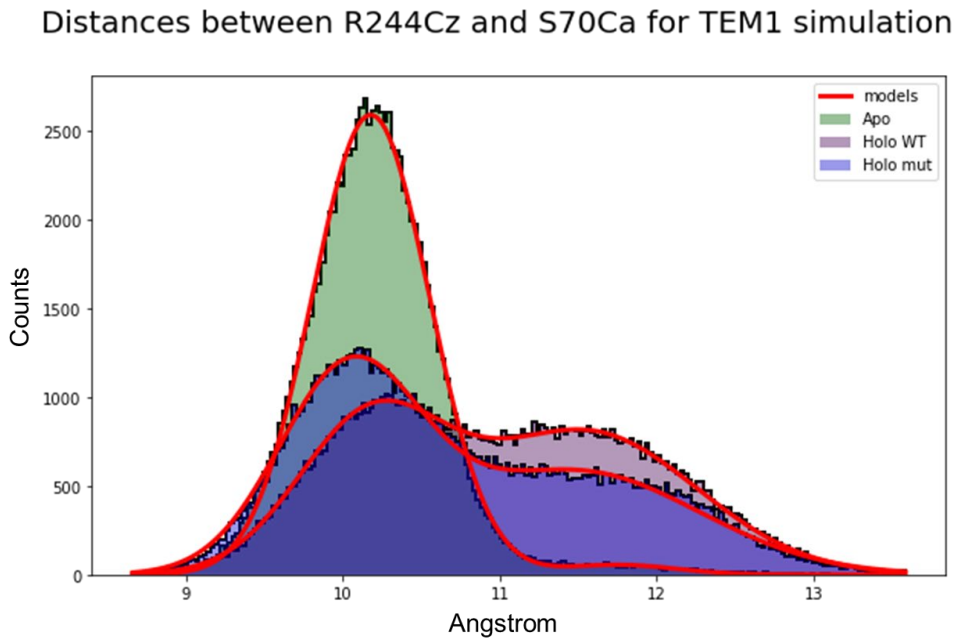


Figure 45. Distance distributions between R244Cζ and S70Cα for the three systems. In red are the calculated bimodal models for each system.

If we only consider the portion of the simulation with the "horizontal" pose, we see only one distribution, centered 11.52 Å away from the pocket.

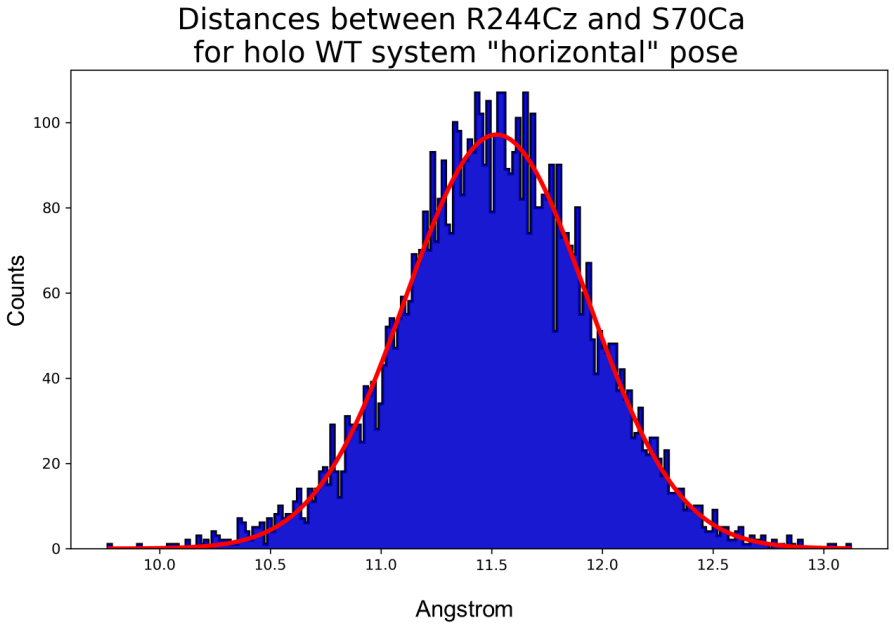


Figure 46. Distance distribution between R244Cζ and S70Cα for the portion of the simulations in the "horizontal" pose. In red is the calculated normal distribution for the system.

4.3.9 Dynamical cross-correlation analysis, betweenness centrality, and average shortest path

Finally, to observe the correlation of motions between residues in the protein, I calculated dynamical cross-correlation (DCC) matrices for the three systems and calculated the differences between the systems. I will only discuss medium and high correlations ($|DCC|_{ij}$ above 0.4) and differences three or more standard deviations larger than the average difference. DCC analysis shows that all the systems have common correlations. All β-sheets are correlated to the neighboring ones, α1 and α12 have correlated motions, α10 is correlated to β3 (residues 230-237). Helices 3, 4, and 5 are all correlated. There are also correlations between α2 and α10 and between α7 and α8. R244 is one of the residues in the β-strands that is most correlated.

When I compare *apo* against *holo* systems, we observe the loss of all anticorrelated motions present in the *apo* system. These anticorrelated motions in the *apo* system are between α1 and β4 and β5, and between α12 and β1 and β2. Most significant changes (three standard deviations or more above the average difference) involve the α/β-domain, the domain that binds FTA (Figure 47). The interdomain correlations seemed to be weakened by FTA, except for the correlations of α11 with α10 and α5. There is an increase in coordinated motions within the α/β-domain, especially between the α1 N-terminus and α12 C-terminus. We also observe an increase in correlations between β3 and β4 and helix α11 and α12. On the residue-specific analysis, we see an increase in correlation in the *holo* mutant between Ser70 and α8. As for R244, we lose the correlation of motions between the arginine and C-terminus of α1 and N-terminus of α12.

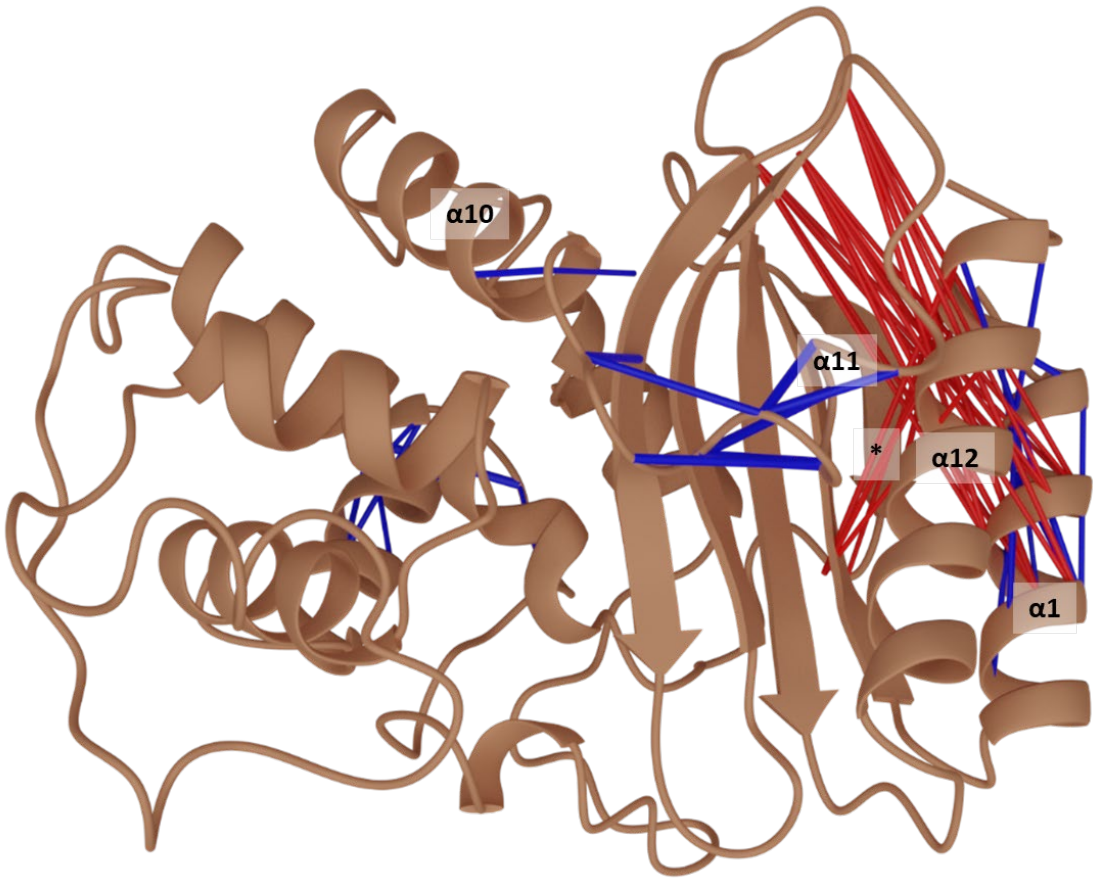


Figure 47. Changes in correlation. In red, we observe a change in anticorrelated motions; in blue, we observe a change in correlated motions. For figure clarity, I only show changes three STD above the average difference and from residues with medium or high correlations. The location where the ligand binds in PDB 1PZP is marked with an asterisk.

To assess how critical a residue is for information transfer within the protein and how FTA impacts that flow of information, I calculated betweenness centrality (BC), which describes how important a residue is to the communication within a protein. As expected, residues at the core of the protein have high BC values. When we introduced FTA, the values of these core residues generally remained unchanged. One notable exception is a product of how BC is calculated; since

we are using Cβ for residues (except for glycines where Cα is used), the gap that FTA forms does artificially decrease the BC for residues in α11 and α12 for the *holo* system.

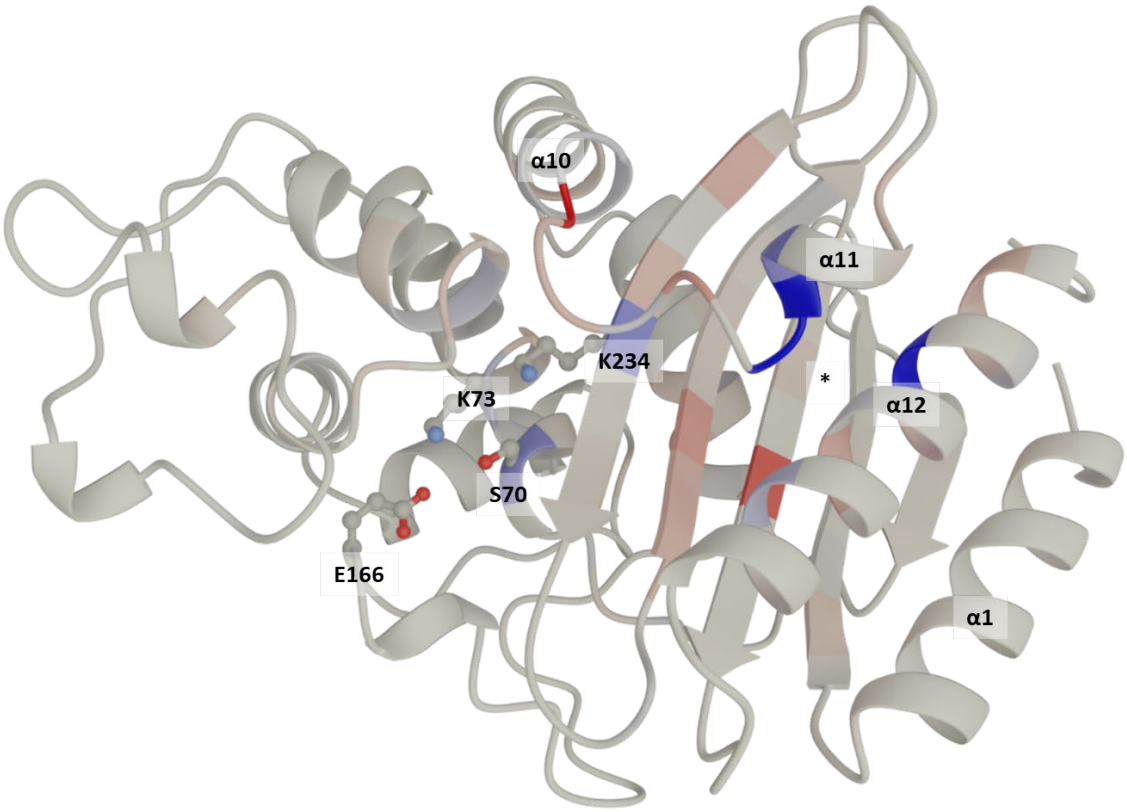


Figure 48. Betweenness centrality (BC) changes between *apo* and *holo* WT simulations. In blue are residues with higher BC in *apo* simulations. In red are residues with higher BC values in *holo* WT simulations. The location where the ligand binds in PDB 1PZP is marked with an asterisk.

It has been shown[210,211] that critical residues tend to have high centralities, so it is interesting that four residues in the catalytic pocket (S70, K73, E166, K234) have lower BC values in the *holo* simulations. This suggests a change in how information (physical or chemical) can pass through the catalytic pocket since residues with high BC control the flow of information. This finding could also explain the impact FTA has on TEM-1 catalysis by altering how the catalytic residues respond to changes in the protein. Another notable difference between *apo* and *holo* simulations is the increase in BC in the *holo* simulations for the N-terminus of β4. This region

includes R244, the residue we have shown to be, on average, further away from the catalytic pocket in the *holo* simulations precluding its ligand stabilization properties.

4.3.10 Allosteric inhibition mechanism by FTA

FTA is an allosteric inhibitor since it binds far away from the catalytic pocket. This ligand changes the global and local dynamics of the whole protein. When it binds to the protein, I observe a rigidification of the protein. Local changes in dynamics are observed in the loop preceding α11, the helix that opens for the cryptic pocket to be formed; this loop is less flexible in the ligand-bound simulations. Especially notable are the changes in side chain dynamics of the residues important for ligand recognition and stabilization: R244 and Y105. R244 in the *apo* simulations is observed mainly pointing towards the orthosteric pocket (seen in the distance analysis), while in the *holo* simulations, the side chain also has a second distribution that has the side chain further away from the catalytic pocket. This second distribution would not be able to stabilize a ligand in the orthosteric pocket, in part explaining the inhibition properties of FTA.

Interestingly I observed a "horizontal" alternative pose in one of the MD simulations. In this pose, the effect observed in R244 is more substantial; all the frames in this pose have R244 pointing away from the orthosteric pocket. This result raises the possibility that the observed "horizontal" conformation is more effective in antibiotic inhibition than the crystallographic pose.

Y105 side chain also presented substantial changes because of FTA binding. In this case, the *apo* simulations presented a conformation more akin to the ones observed in simulations of the benzylpenicillin/TEM-1 complex.[212] While the *holo* simulations had the side chain packed against P107, in this side chain conformation, the phenol group of Y105 will not be able to interact with

the antibiotic's aromatic groups, which would further destabilize the binding of an antibiotic molecule in the orthosteric pocket.

I also observed how the BC of critical residues changes in the simulations. S70, one of the catalytic residues, has a higher BC value in the *apo* simulations. This decrease in BC for the *holo* simulations suggests that ligand binding reduces the influence on inter-protein communication of S70. Another impacted residue is R244; this residue's influence on information transmission is increased in the *holo* simulations. These changes suggest how FTA binding alters how the protein responds to external changes, reducing how important residues for catalysis can react to changes in the protein (*i.e.*, antibiotic binding) and increasing the response of one residue that has been shown to diminish ligand binding.

4.4 Conclusions

In this work, I used molecular dynamics simulations and network analysis of TEM-1 with and without the allosteric inhibitor FTA to propose a mechanism by which the small molecule inhibits TEM-1. This inhibitor lies ≈15 Å away from the orthosteric pocket. Ligand binding has profound effects on protein local and global dynamics; even though TEM-1 is a rigid protein, the binding of FTA increases the rigidity of the protein even more. This rigidity increase was confirmed by RMSD, RMSF, and PCA analysis and agreed with studies of TEM-1 binding to the protein inhibitor BLIP.[199,200] The allosteric ligand also has important effects on the population shifts of side chain conformations of critical residues. Notably, R244 has a significant population change. In the *apo* simulations, it assumes a single conformation and lies a consistent distance from the pocket. In the *holo* simulations, the distance has a binary distribution, and the residue

assumes a different main conformation per Janin plots. The ligand also changes the protein network and the communication between residues, as seen by the changes in correlated motions, BC, and L. These changes could influence how catalytic residues respond to binding in the orthosteric pocket, likely leading to changes in catalysis.

Here, I also show a novel binding mode for FTA. As support for the existence of the "horizontal" pose could be rationalized because of the presence of a similar binding pocket, as observed in the structure of N,N-bis(4-chlorobenzyl)-1H-1,2,3,4-tetrazole-5-amine bound TEM-1 (PDB 1PZO). Crystallographic contacts of tetrazole moiety in FTA could stabilize the crystallographic pose, which has FTA partially embedded in the cryptic pocket. This new insight can help medicinal chemists explore a different pocket conformation to target β-lactamases as they search for new molecules to combat multi-resistance strains. To conclude, to our knowledge, our study is the first to propose a "horizontal" pose for FTA binding, helping medicinal chemists in their quest to combat MDR strains.

5.0 Conclusions and future directions

5.1 Chapter 1 - Introduction to computer-aided drug design

In chapter one, I briefly introduced the field of computer-aided drug design (CADD). Here I talked about the importance of CADD and how it has advanced drug discovery. I introduced some of the structure-based drug-design techniques I used in this dissertation.

One of the biggest challenges in drug discovery and in CADD is the low hit rates observed in lead discovery, and ensemble docking is helping address this problem.[213] In this dissertation, I described the development of a tool that leads to better protein-pocket sampling, helping better incorporate protein flexibility into docking, which will then help increase the hit rates in docking algorithms.

In the next two chapters, I used molecular dynamics simulations to understand how a perturbation (mutation or ligand binding) affects protein dynamics. I used several analysis tools to uncover the mechanisms by which a ligand or a mutation inhibits catalysis. I also undertook a more traditional CADD project that uses ensemble docking to identify candidate inhibitors for experimental testing.

5.2 Chapter 2 – Sub-Pocket EXplorer (SubPEx): Leveraging weighted ensemble simulations to enhance the conformational sampling of binding-pocket conformations

In chapter two, I described how I developed Sub-Pocket EXplorer (SubPEx). This tool uses weighted ensemble path sampling to accelerate the sampling of protein pocket conformations for later use in ensemble docking. I showed that the composite RMSD, a linear combination of pocket and backbone RMSD, outperformed all other progress coordinates as well as vanilla MD simulations. I also demonstrated how clustering by generation improves the clustering speed and the diversity of pocket conformations. Finally, I tested SubPEx on biologically relevant proteins involved in cancer or viral infections.

Before publishing the SubPEx manuscript, I need to repeat the simulations for neuraminidase starting from its closed conformation. As mentioned in the chapter, the system had not fully equilibrated because the crystal structure has a ligand in the pocket, stabilizing the conformation observed in the experimentally determined protein structure. The system needs more time to equilibrate correctly.

My current implementation ignores the probabilities that WESTPA calculates. Since we are nowhere near the equilibrium population distribution, we cannot use the WESTPA probabilities in subsequent analysis without reweighting them so they are closer to the actual equilibrium probabilities. I propose using the history augmented Markov state models plugin developed for WESTPA[48], or a Markov state model approach.

Once the walkers are reweighted, we could use the walker probabilities we obtain from WE to calculate the probabilities of the clusters obtained by the clustering algorithm. These cluster probabilities could then be assigned to the centroids for use in Boltzmann docking, an ensemble

docking technique that weights the predicted binding affinities (scores) of each state by its equilibrium probability.[214,215]

5.3 Chapter 3 – Novel mutation in hexokinase 2 confers resistance to 2-deoxyglucose by altering protein dynamics

In chapter three, I showed how molecular dynamics simulations could explain a mutation's role in the resistance mechanism of a known inhibitor. This published work [90] shows the power of combining experimental and computational tools to explain a phenomenon at atomistic resolution.

We demonstrated how a change in a single residue in a 469-residue protein can greatly impact catalysis, even when that residue does not interact with the substrate. We studied hexokinase II, a protein relevant to the catabolism of glucose. A mutation in this protein allows yeast cells to grow in the presence of 2-deoxy-glucose, an environment typically not conducive to life. The G238V mutation changed the dynamics of the whole protein, impacting it in large conformational changes as well as in small, more localized changes. Using network analysis applied to the molecular dynamics simulations, I observed how the mutation changes the flow of intra-protein communication.

The mutation substantially impacted residue V236, and I hypothesized that V236 plays a crucial role in determining how the protein transitions from the open to the closed conformation. The impact of V236 on global dynamics could be tested experimentally using mutagenesis. I hypothesize that a larger side chain (e.g., isoleucine, leucine, to maintain the hydrophobicity) or one that cannot rotate (e.g., proline) will further hinder the open-to-close transition. In contrast, glycine, a more flexible residue, should enable the transition.

Another hypothesis that requires experimental validation is the proposed increase in the entropic penalty for glucose binding. I hypothesized this penalty would be more significant for the mutated protein because we observed higher fluctuations in the binding cleft. This hypothesis could be proven correct by isothermal titration calorimetry (ITC), which determines binding affinities and the enthalpic and entropic contributions to binding.

5.4 Chapter 4 – Mentoring undergraduates in computational research

In chapter four, I showed results I obtained with undergraduate students' help. I introduced the chapter by describing the importance of undergraduate research experiences, their impact on the students, and the future of STEM fields.

In the first half of the chapter, we used a traditional computer-aided drug design approach to lead discovery. We first generated six models of the human protein SMUG1, a protein involved in base excision repair. We next docked three different small molecule databases into the six protein conformations.

We concluded this project by recommending our collaborators test some of these small molecules for experimental validation. The next step will be to optimize any compound that binds the human protein using structure-activity relationship studies.

In the second project, I used molecular dynamics simulations of a protein-ligand complex to understand how the ligand influences the protein. The system we studied consists of an allosteric small-molecule inhibitor that binds TEM-1 lactamase, one of the proteins responsible for antibiotic resistance in bacteria. Here I performed molecular dynamics simulations of the protein bound to the inhibitor and tracked the changes in protein dynamics observed due to ligand binding.

One of the most exciting results is that the ligand inserted itself further into the allosteric pocket. This insertion only happened in one of our WT TEM-1 simulations, but I pursued this unexpected ligand pose to understand the alternate pose's viability. As a follow-up experiment, I would perform NMR experiments, specifically chemical shift perturbation (CSP) analysis.[216] CSP has been extensively used to track the location of ligand binding by analyzing changes in protein chemical shifts due to ligand binding.

Dihedral-angle and network analysis suggest a mechanism by which the small molecule inhibits catalytic activity. This study can help medicinal chemists develop new small molecule inhibitors that bind the allosteric pocket, advancing the fight against multidrug-resistant bacteria.

Bibliography

1. Macalino, S. J. Y., Gosu, V., Hong, S. & Choi, S. Role of computer-aided drug design in modern drug discovery. *Arch Pharm Res* **38**, 1686–1701 (2015).
2. Schlander, M., Hernandez-Villafuerte, K., Cheng, C.-Y., Mestre-Ferrandiz, J. & Baumann, M. How Much Does It Cost to Research and Develop a New Drug? A Systematic Review and Assessment. *Pharmacoeconomics* **39**, 1243–1269 (2021).
3. Wong, C. H., Siah, K. W. & Lo, A. W. Estimation of clinical trial success rates and related parameters. *Biostatistics* **20**, 273–286 (2019).
4. Jorgensen, W. L. The Many Roles of Computation in Drug Discovery. *Science (1979)* **303**, 1813–1818 (2004).
5. Greer, J., Erickson, J. W., Baldwin, J. J. & Varney, M. D. Application of the Three-Dimensional Structures of Protein Target Molecules in Structure-Based Drug Design. *J Med Chem* **37**, 1035–1054 (1994).
6. Clark, D. E. What has computer-aided molecular design ever done for drug discovery? *Expert Opin Drug Discov* **1**, 103–110 (2006).
7. Leelananda, S. P. & Lindert, S. Computational methods in drug discovery. *Beilstein Journal of Organic Chemistry* **12**, 2694–2718 (2016).
8. Illergård, K., Ardell, D. H. & Elofsson, A. Structure is three to ten times more conserved than sequence—A study of structural response in protein cores. *Proteins: Structure, Function, and Bioinformatics* **77**, 499–508 (2009).
9. Muhammed, M. T. & Aki-Yalcin, E. Homology modeling in drug discovery: Overview, current applications, and future perspectives. *Chem Biol Drug Des* **93**, 12–20 (2019).

10. Cavasotto, C. N. & Phatak, S. S. Homology modeling in drug discovery: current trends and applications. *Drug Discov Today* **14**, 676–683 (2009).

11. Schauperl, M. & Denny, R. A. AI-Based Protein Structure Prediction in Drug Discovery: Impacts and Challenges. *J Chem Inf Model* (2022) doi:10.1021/ACS.JCIM.2C00026.

12. Hameduh, T., Haddad, Y., Adam, V. & Heger, Z. Homology modeling in the time of collective and artificial intelligence. *Comput Struct Biotechnol J* **18**, 3494–3506 (2020).

13. Ma, J., Wang, S., Zhao, F. & Xu, J. Protein threading using context-specific alignment potential. *Bioinformatics* **29**, i257–i265 (2013).

14. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature 2021 596:7873* **596**, 583–589 (2021).

15. Hillisch, A., Pineda, L. F. & Hilgenfeld, R. Utility of homology models in the drug discovery process. *Drug Discov Today* **9**, 659–669 (2004).

16. Hertzberg, R. P. & Pope, A. J. High-throughput screening: new technology for the 21st century. *Curr Opin Chem Biol* **4**, 445–451 (2000).

17. Volochnyuk, D. M. *et al.* Evolution of commercially available compounds for HTS. *Drug Discov Today* **24**, 390–402 (2019).

18. Liu, H. & Begley, T. *Comprehensive Natural Products III*. (Elsevier Science, 2020).

19. Liu, R., Li, X. & Lam, K. S. Combinatorial chemistry in drug discovery. *Curr Opin Chem Biol* **38**, 117–126 (2017).

20. A. Srinivas Reddy, S. Priyadarshini Pati, P. Praveen Kumar, H.N. Pradeep & G. Narahari Sastry. Virtual Screening in Drug Discovery - A Computational Perspective. *Curr Protein Pept Sci* **8**, 329–351 (2007).

21. Taylor, R. D., Jewsbury, P. J. & Essex, J. W. A review of protein-small molecule docking methods. *Journal of Computer-Aided Molecular Design 2002 16:3* **16**, 151–166 (2002).

22. Pagadala, N. S., Syed, K. & Tuszynski, J. Software for molecular docking: a review. *Biophys Rev* **9**, 91–102 (2017).

23. Bentham Science Publisher, B. S. P. Scoring Functions for Protein-Ligand Docking. *Curr Protein Pept Sci* **7**, 407–420 (2006).

24. Amaro, R. E., Baron, R. & McCammon, J. A. An improved relaxed complex scheme for receptor flexibility in computer-aided drug design. *J Comput Aided Mol Des* **22**, 693–705 (2008).

25. Amaro, R. E. *et al.* Ensemble Docking in Drug Discovery. *Biophys J* **114**, 2271–2278 (2018).

26. Bahar, I., Jernigan, R. L. & Dill, K. A. *Protein Actions: Principles and Modeling*. (Garland Science, 2017).

27. Leach, A. *Molecular Modelling: Principles and Applications*. (Pearson, 2001).

28. Durrant, J. D. & McCammon, J. A. Molecular dynamics simulations and drug discovery. *BMC Biol* **9**, 1–9 (2011).

29. Liu, K., Watanabe, E. & Kokubo, H. Exploring the stability of ligand binding modes to proteins by molecular dynamics simulations. *J Comput Aided Mol Des* **31**, 201–211 (2017).

30. Liu, K. & Kokubo, H. Exploring the Stability of Ligand Binding Modes to Proteins by Molecular Dynamics Simulations: A Cross-docking Study. *J Chem Inf Model* **57**, 2514–2522 (2017).

31. Smith, R. D. & Carlson, H. A. Identification of Cryptic Binding Sites Using MixMD with Standard and Accelerated Molecular Dynamics. *J Chem Inf Model* **61**, 1287–1299 (2021).

32. Vajda, S., Beglov, D., Wakefield, A. E., Egbert, M. & Whitty, A. Cryptic binding sites on proteins: definition, detection, and druggability. *Curr Opin Chem Biol* **44**, 1–8 (2018).

33. Kuzmanic, A., Bowman, G. R., Juarez-Jimenez, J., Michel, J. & Gervasio, F. L. Investigating Cryptic Binding Sites by Molecular Dynamics Simulations. *ACS Appl Mater Interfaces* (2020) doi:10.1021/acs.accounts.9b00613

34. Cournia, Z., Allen, B. & Sherman, W. Relative Binding Free Energy Calculations in Drug Discovery: Recent Advances and Practical Considerations. *J Chem Inf Model* **57**, 2911–2937 (2017).

35. Polishchuk, P. Interpretation of Quantitative Structure-Activity Relationship Models: Past, Present, and Future. *J Chem Inf Model* **57**, 2618–2639 (2017).

36. Frye, L., Bhat, S., Akinsanya, K. & Abel, R. From computer-aided drug discovery to computer-driven drug discovery. *Drug Discov Today Technol* **39**, 111–117 (2021).

37. Burki, T. A new paradigm for drug development. *Lancet Digit Health* **2**, e226–e227 (2020).

38. Ricci-Lopez, J., Aguila, S. A., Gilson, M. K. & Brizuela, C. A. Improving Structure-Based Virtual Screening with Ensemble Docking and Machine Learning. *J Chem Inf Model* **61**, 5362–5376 (2021).

39. Acharya, A. *et al.* Supercomputer-Based Ensemble Docking Drug Discovery Pipeline with Application to Covid-19. *J Chem Inf Model* **60**, 5832–5852 (2020).

40. van Meel, J. A., Arnold, A., Frenkel, D., Portegies Zwart, S. F. & Belleman, R. G. Harvesting graphics power for MD simulations. *https://doi.org/10.1080/08927020701744295* **34**, 259–266 (2008).

41. Glaser, J. *et al.* Strong scaling of general-purpose molecular dynamics simulations on GPUs. *Comput Phys Commun* **192**, 97–107 (2015).

42. E, W., Ren, W. & Vanden-Eijnden, E. String method for the study of rare events. *Phys Rev B Condens Matter Mater Phys* **66**, 523011–523014 (2002).

43. Chong, L. T., Saglam, A. S. & Zuckerman, D. M. Path-sampling strategies for simulating rare events in biomolecular systems. *Curr Opin Struct Biol* **43**, 88–94 (2017).

44. Dellago, C. & Bolhuis, P. G. Transition Path Sampling and Other Advanced Simulation Techniques for Rare Events. in *Advanced Computer Simulation Approaches for Soft Matter Sciences III* 167–233 (Springer Berlin Heidelberg, 2009). doi:10.1007/978-3-540-87706-6_3.

45. Zuckerman, D. M. & Chong, L. T. Weighted Ensemble Simulation: Review of Methodology, Applications, and Software. *Annu Rev Biophys* **46**, 43–57 (2017).

46. Huber, G. A. & Kim, S. Weighted-ensemble Brownian dynamics simulations for protein association reactions. *Biophys J* **70**, 97–110 (1996).

47. Zwier, M. C. *et al.* WESTPA: An interoperable, highly scalable software package for weighted ensemble simulation and analysis. *J Chem Theory Comput* **11**, 800–809 (2015).

48. Russo, J. D. *et al.* WESTPA 2.0: High-Performance Upgrades for Weighted Ensemble Simulations and Analysis of Longer-Timescale Applications. *J Chem Theory Comput* **18**, 638–649 (2022).

49. Burley, S. K. *et al.* RCSB Protein Data Bank: powerful new tools for exploring 3D structures of biological macromolecules for basic and applied research and education in fundamental biology, biomedicine, biotechnology, bioengineering and energy sciences. *Nucleic Acids Res* **49**, D437–D451 (2021).

50. Amaral, M. *et al.* Protein conformational flexibility modulates kinetics and thermodynamics of drug binding. *Nature Communications 2017 8:1* **8**, 1–14 (2017).

51. Russell, R. J. *et al.* The structure of H5N1 avian influenza neuraminidase suggests new opportunities for drug design. *Nature 2006 443:7107* **443**, 45–49 (2006).

52. Kuser, P. R., Krauchenco, S., Antunes, O. A. C. & Polikarpov, I. The High Resolution Crystal Structure of Yeast Hexokinase PII with the Correct Primary Sequence Provides New Insights into Its Mechanism of Action *. *Journal of Biological Chemistry* **275**, 20814–20821 (2000).

53. Jurrus, E. *et al.* Improvements to the APBS biomolecular solvation software suite. *Protein Science* **27**, 112–128 (2018).

54. Li, H., Robertson, A. D. & Jensen, J. H. Very fast empirical prediction and rationalization of protein pKa values. *Proteins: Structure, Function, and Bioinformatics* **61**, 704–721 (2005).

55. D.A. Case, K. Belfon, I.Y. Ben-Shalom, S.R. Brozell, D.S. Cerutti, T.E. Cheatham, III, V. W. D. C. *et al.* Amber 2020. *Journal of Chemical Information and Modeling* vol. 53 1689–1699 Preprint at https://ambermd.org/ (2020).

56. Maier, J. A. *et al.* ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *J Chem Theory Comput* **11**, 3696–3713 (2015).

57. Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W. & Klein, M. L. Comparison of simple potential functions for simulating liquid water. *J Chem Phys* **79**, 926–935 (1983).

58. Phillips, J. C. *et al.* Scalable molecular dynamics on CPU and GPU architectures with NAMD. *Journal of Chemical Physics* **153**, 44130 (2020).

59. Torrillo, P. A., Bogetti, A. T. & Chong, L. T. A minimal, adaptive binning scheme for weighted ensemble simulations. *Journal of Physical Chemistry A* **125**, 1642–1649 (2021).

60. Michaud-Agrawal, N., Denning, E. J., Woolf, T. B. & Beckstein, O. MDAnalysis: A toolkit for the analysis of molecular dynamics simulations. *J Comput Chem* **32**, 2319–2327 (2011).

61. Gowers, R. *et al.* MDAnalysis: A Python Package for the Rapid Analysis of Molecular Dynamics Simulations. in *Proceedings of the 15th Python in Science Conference* (eds. Benthall, S. & Rostrup, S.) 98–105 (2016). doi:10.25080/Majora-629e541a-00e.

62. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **12**, 2825–2830 (2011).

63. Roe, D. R. & Cheatham, T. E. PTRAJ and CPPTRAJ: Software for processing and analysis of molecular dynamics trajectory data. *J Chem Theory Comput* **9**, 3084–3095 (2013).

64. Shao, J., Tanner, S. W., Thompson, N. & Cheatham, T. E. Clustering molecular dynamics trajectories: 1. Characterizing the performance of different clustering algorithms. *J Chem Theory Comput* **3**, 2312–2334 (2007).

65. Durrant, J. D., de Oliveira, C. A. F. & McCammon, J. A. POVME: An algorithm for measuring binding-pocket volumes. *J Mol Graph Model* **29**, 773–776 (2011).

66. Durrant, J. D., Votapka, L., Sørensen, J. & Amaro, R. E. POVME 2.0: An enhanced tool for determining pocket shape and volume characteristics. *J Chem Theory Comput* **10**, 5047–5056 (2014).

67. Sztain, T. *et al.* A glycan gate controls opening of the SARS-CoV-2 spike protein. *Nature Chemistry 2021 13:10* **13**, 963–968 (2021).

68. Lotz, S. D. & Dickson, A. Unbiased Molecular Dynamics of 11 min Timescale Drug Unbinding Reveals Transition State Stabilizing Interactions. *J Am Chem Soc* **140**, 618–628 (2018).

69. Dickson, A., Mustoe, A. M., Salmon, L. & Brooks, C. L. Efficient in silico exploration of RNA interhelical conformations using Euler angles and WExplore. *Nucleic Acids Res* **42**, 12126–12137 (2014).

70. Birbo, B., Madu, E. E., Madu, C. O., Jain, A. & Lu, Y. Role of HSP90 in Cancer. *Int J Mol Sci* **22**, 10317 (2021).

71. Mahalingam, D. *et al.* Targeting HSP90 for cancer therapy. *Br J Cancer* **100**, 1523–1529 (2009).

72. Schopf, F. H., Biebl, M. M. & Buchner, J. The HSP90 chaperone machinery. *Nat Rev Mol Cell Biol* **18**, 345–360 (2017).

73. Jackson, S. E. Hsp90: Structure and Function. in *Molecular Chaperones* vol. 328 155–240 (Springer, Berlin, Heidelberg, 2012).

74. Influenza (Seasonal). https://www.who.int/en/news-room/fact-sheets/detail/influenza-(seasonal).

75. Moscona, A. Neuraminidase Inhibitors for Influenza. *https://doi.org/10.1056/NEJMra050740* **353**, 1363–1373 (2005).

76. Colman, P. M. Neuraminidase. *The Influenza Viruses* 175–218 (1989) doi:10.1007/978-1-4613-0811-9_4.

77. Gasparini, R., Amicizia, D., Lai, P. L., Bragazzi, N. L. & Panatto, D. Compounds with anti-influenza activity: present and future of strategies for the optimal treatment and management of influenza. Part I: Influenza life-cycle and currently available drugs. *J Prev Med Hyg* **55**, 69–85 (2014).

78. Wang, M. *et al.* Influenza A Virus N5 Neuraminidase Has an Extended 150-Cavity. *J Virol* **85**, 8431–8435 (2011).

79. McAuley, J. L., Gilbertson, B. P., Trifkovic, S., Brown, L. E. & McKimm-Breschkin, J. L. Influenza virus neuraminidase structure and functions. *Front Microbiol* **10**, 39 (2019).

80. Zima, V. *et al.* Investigation of flexibility of neuraminidase 150-loop using tamiflu derivatives in influenza A viruses H1N1 and H5N1. *Bioorg Med Chem* **27**, 2935–2947 (2019).

81. Amaro, R. E. *et al.* Mechanism of 150-cavity formation in influenza neuraminidase. *Nature Communications 2011 2:1* **2**, 1–7 (2011).

82. Ciscato, F., Ferrone, L., Masgras, I., Laquatra, C. & Rasola, A. Hexokinase 2 in Cancer: A Prima Donna Playing Multiple Characters. *International Journal of Molecular Sciences 2021, Vol. 22, Page 4716* **22**, 4716 (2021).

83. Slein, M. W., Cori, G. T. & Cori, C. F. A comparative study of hexokinase from yeast and animal tissues. *J Biol Chem* **186**, 763–80 (1950).

84. Liberti, M. v. & Locasale, J. W. The Warburg Effect: How Does it Benefit Cancer Cells? *Trends in Biochemical Sciences* vol. 41 211–218 Preprint at https://doi.org/10.1016/j.tibs.2015.12.001 (2016).

85. Mathupala, S. P., Ko, Y. H. & Pedersen, P. L. The pivotal roles of mitochondria in cancer: Warburg and beyond and encouraging prospects for effective therapies. *Biochimica et Biophysica Acta - Bioenergetics* vol. 1797 1225–1230 Preprint at https://doi.org/10.1016/j.bbabio.2010.03.025 (2010).

86. Feron, O. Pyruvate into lactate and back: From the Warburg effect to symbiotic energy fuel exchange in cancer cells. *Radiotherapy and Oncology* vol. 92 329–333 Preprint at https://doi.org/10.1016/j.radonc.2009.06.025 (2009).

87. Kuettner, E. B. *et al.* Crystal structure of hexokinase KlHxk1 of kluyveromyces lactis: A molecular basis for understanding the control of yeast hexokinase functions via covalent modification and oligomerization. *Journal of Biological Chemistry* **285**, 41019–41033 (2010).

88. Chodera, J. D. A Simple Method for Automated Equilibration Detection in Molecular Simulations. *J Chem Theory Comput* **12**, 1799–1805 (2016).

89. Smith, L. J., Daura, X. & van Gunsteren, W. F. Assessing equilibration and convergence in biomolecular simulations. *Proteins: Structure, Function, and Bioinformatics* **48**, 487–496 (2002).

90. Hellemann, E. *et al.* Novel mutation in hexokinase 2 confers resistance to 2-deoxyglucose by altering protein dynamics. *PLoS Comput Biol* **18**, e1009929 (2022).

91. Pastorino, J. & Hoek, J. Hexokinase II: The Integration of Energy Metabolism and Control of Apoptosis. *Curr Med Chem* **10**, 1535–1551 (2003).

92. Mathupala, S. P., Ko, Y. H. & Pedersen, P. L. Hexokinase-2 bound to mitochondria: Cancer's stygian link to the "Warburg effect" and a pivotal target for effective therapy. *Semin Cancer Biol* **19**, 17–24 (2009).

93. Warburg, O. H. *The metabolism of tumours: investigations from the Kaiser Wilhelm Institute for Biology, Berlin-Dahlem*. (Constable & Company Limited, 1930).

94. Pastorino, J. G., Shulga, N. & Hoek, J. B. Mitochondrial binding of hexokinase II inhibits Bax-induced cytochrome c release and apoptosis. *Journal of Biological Chemistry* **277**, 7610–7618 (2002).

95. Kuser, P. R., Krauchenco, S., Antunes, O. A. C. & Polikarpov, I. The high resolution crystal structure of yeast hexokinase PII with the correct primary sequence provides new insights into its mechanism of action. *Journal of Biological Chemistry* **275**, 20814–20821 (2000).

96. Kuser, P., Cupri, F., Bleicher, L. & Polikarpov, I. Crystal structure of yeast hexokinase PI in complex with glucose: A classical "induced fit" example revised. *Proteins: Structure, Function, and Bioinformatics* **72**, 731–740 (2008).

97. Jeong, E. J. *et al.* Detection of glucose-induced conformational change in hexokinase II using fluorescence complementation assay. *Biotechnol Lett* **29**, 797–802 (2007).

98. Shoham, M. & Steitz, T. A. The 6-hydroxymethyl group of a hexose is essential for the substrateinduced closure of the cleft in hexokinase. *Biochimica et Biophysica Acta (BBA)/Protein Structure and Molecular* **705**, 380–384 (1982).

99. McCartney, R. R., Chandrashekarappa, D. G., Zhang, B. B. & Schmidt, M. C. Genetic Analysis of Resistance and Sensitivity to 2-Deoxyglucose in Saccharomyces cerevisiae. *Genetics* **198**, 635–646 (2014).

100. Biely, P., Krátký, Z., Kovarík, J. & Bauer, S. Effect of 2-Deoxyglucose on Cell Wall Formation in Saccharomyces cerevisiae and Its Relation to Cell Growth Inhibition. *J Bacteriol* **107**, 121–129 (1971).

101. Soncini, S. R. *et al.* Spontaneous mutations that confer resistance to 2-deoxyglucose act through Hxk2 and Snf1 pathways to regulate gene expression and HXT endocytosis. *PLoS Genet* **16**, e1008484 (2020).

102. Kang, H. T. & Hwang, E. S. 2-Deoxyglucose: An anticancer and antiviral therapeutic, but not any more a low glucose mimetic. *Life Sci* **78**, 1392–1399 (2006).

103. Dwarakarnath, B. S. & Jain, V. Targeting glucose metabolism with 2-deoxy-D-glucose for improving cancer therapy. *http://dx.doi.org/10.2217/fon.09.44* **5**, 581–585 (2009).

104. Granchi, C., Fancelli, D. & Minutolo, F. An update on therapeutic opportunities offered by cancer glycolytic metabolism. *Bioorg Med Chem Lett* **24**, 4915–4925 (2014).

105. Dwarakanath, B. S. *et al.* Clinical studies for improving radiotherapy with 2-deoxy-D-glucose: Present status and future prospects. *J Cancer Res Ther* **5**, 21 (2009).

106. O'Donnell, A. F., Huang, L., Thorner, J. & Cyert, M. S. A calcineurin-dependent switch controls the trafficking function of α-arrestin Aly1/Art6. *Journal of Biological Chemistry* **288**, 24063–24080 (2013).

107. Gietz, R. D. & Woods, R. A. Transformation of yeast by lithium acetate/single-stranded carrier DNA/polyethylene glycol method. *Methods Enzymol* **350**, 87–96 (2002).

108. Ottilie, S. *et al.* Rapid Chagas Disease Drug Target Discovery Using Directed Evolution in Drug-Sensitive Yeast. *ACS Chem Biol* **12**, 422–434 (2017).

109. Ottilie, S. *et al.* Two inhibitors of yeast plasma membrane ATPase 1 (ScPma1p): Toward the development of novel antifungal therapies. *J Cheminform* **10**, 1–9 (2018).

110. Suzuki, Y. *et al.* The green monster process for the generation of yeast strains carrying multiple gene deletions. *Journal of Visualized Experiments* (2012) doi:10.3791/4072.

111. Ottilie, S. *et al.* Adaptive laboratory evolution in S. cerevisiae highlights role of transcription factors in fungal xenobiotic resistance. *Communications Biology 2022 5:1* **5**, 1–14 (2022).

112. Hoffman, C. S. & Winston, F. A ten-minute DNA preparation from yeast efficiently releases autonomous plasmids for transformaion of Escherichia coli. *Gene* **57**, 267–272 (1987).

113. Soncini, S. R. *et al.* Spontaneous mutations that confer resistance to 2-deoxyglucose act through Hxk2 and Snf1 pathways to regulate gene expression and HXT endocytosis. *PLoS Genet* **16**, e1008484 (2020).

114. Ellison, M. A., Walker, J. L., Ropp, P. J., Durrant, J. D. & Arndt, K. M. MutantHuntWGS: A Pipeline for Identifying Saccharomyces cerevisiae Mutations. *G3 Genes|Genomes|Genetics* **10**, 3009–3014 (2020).

115. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nature Methods 2012 9:4* **9**, 357–359 (2012).

116. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).

117. Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011).

118. Cingolani, P. *et al.* A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly (Austin)* **6**, 80–92 (2012).

119. Minear, S. *et al.* Curcumin inhibits growth of saccharomyces cerevisiae through iron chelation. *Eukaryot Cell* **10**, 1574–1581 (2011).

120. Volland, C., Urban-Grimal, D., Géraud, G. & Haguenauer-Tsapis, R. Endocytosis and degradation of the yeast uracil permease under adverse conditions. *Journal of Biological Chemistry* **269**, 9833–9841 (1994).

121. Goldstein, A. & Oliver Lampen, J. [76] β-d-Fructofuranoside fructohydrolase from yeast. *Methods Enzymol* **42**, 504–511 (1975).

122. The PyMOL Molecular Graphics System.

123. Anderson, C. M., Stenkamp, R. E. & Steitz, T. A. Sequencing a protein by X-ray crystallography: II. Refinement of yeast hexokinase B Co-ordinates and sequence at 2.1 Å resolution. *J Mol Biol* **123**, 15–33 (1978).

124. Nawaz, M. H. *et al.* The catalytic inactivation of the N-half of human hexokinase 2 and structural and biochemical characterization of its mitochondrial conformation. *Biosci Rep* **38**, (2018).

125. Kirschner, K. N. *et al.* GLYCAM06: A generalizable biomolecular force field. Carbohydrates. *J Comput Chem* **29**, 622–655 (2008).

126. Phillips, J. C. *et al.* Scalable molecular dynamics with NAMD. *J Comput Chem* **26**, 1781–1802 (2005).

127. Ryckaert, J. P., Ciccotti, G. & Berendsen, H. J. C. Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *J Comput Phys* **23**, 327–341 (1977).

128. Brown, D. K. *et al.* MD-TASK: a software suite for analyzing molecular dynamics trajectories. *Bioinformatics* **33**, 2768–2771 (2017).

129. Ottilie, S. *et al.* Rapid Chagas Disease Drug Target Discovery Using Directed Evolution in Drug-Sensitive Yeast. *ACS Chem Biol* **12**, 422–434 (2017).

130. Luth, M. R., Gupta, P., Ottilie, S. & Winzeler, E. A. Using in Vitro Evolution and Whole Genome Analysis to Discover Next Generation Targets for Antimalarial Drug Discovery. *ACS Infect Dis* **4**, 301–314 (2018).

131. Goldgof, G. M. *et al.* Comparative chemical genomics reveal that the spiroindolone antimalarial KAE609 (Cipargamin) is a P-type ATPase inhibitor. *Scientific Reports 2016 6:1* **6**, 1–13 (2016).

132. Ottilie, S. *et al.* Two inhibitors of yeast plasma membrane ATPase 1 (ScPma1p): Toward the development of novel antifungal therapies. *J Cheminform* **10**, 1–9 (2018).

133. Suzuki, Y. *et al.* The Green Monster Process for the Generation of Yeast Strains Carrying Multiple Gene Deletions. *JoVE (Journal of Visualized Experiments)* e4072 (2012) doi:10.3791/4072.

134. Suzuki, Y. *et al.* Knocking out multigene redundancies via cycles of sexual assortment and fluorescence selection. *Nature Methods 2011 8:2* **8**, 159–164 (2011).

135. Altschul, S. F. & Gish, W. [27] Local alignment statistics. *Methods Enzymol* **266**, 460–480 (1996).

136. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J Mol Biol* **215**, 403–410 (1990).

137. Defenouillère, Q. *et al.* The induction of HAD-like phosphatases by multiple signaling pathways confers resistance to the metabolic inhibitor 2-deoxyglucose. *Sci Signal* **12**, (2019).

138. Aft, R. L., Zhang, F. W. & Gius, D. Evaluation of 2-deoxy-D-glucose as a chemotherapeutic agent: mechanism of cell death. *Br J Cancer* **87**, 805–812 (2002).

139. Zhang, D. *et al.* 2-Deoxy-D-glucose targeting of glucose metabolism in cancer cells as a potential therapy. *Cancer Lett* **355**, 176–183 (2014).

140. Linn, M. C., Palmer, E., Baranger, A., Gerard, E. & Stone, E. Undergraduate research experiences: Impacts and opportunities. *Science (1979)* **347**, (2015).

141. Rodríguez Amaya, L., Betancourt, T., Collins, K. H., Hinojosa, O. & Corona, C. Undergraduate research experiences: mentoring, awareness, and perceptions—a case study at a Hispanic-serving institution. *Int J STEM Educ* **5**, 1–13 (2018).

142. Vincent-Ruz, P., Grabowski, J. & Schunn, C. D. The Impact of Early Participation in Undergraduate Research Experiences on Multiple Measures of Premed Path Success. *Counc Undergrad Res Q* **1**, 13–18 (2018).

143. Russell, S. H., Hancock, M. P. & McCullough, J. Benefits of undergraduate research experiences. *Science (1979)* **316**, 548–549 (2007).

144. of Sciences, E. M. *et al. Undergraduate Research Experiences for STEM Students*. (National Academies Press, 2017). doi:10.17226/24622.

145. Martin, L. J. DNA Damage and RepairRelevance to Mechanisms of Neurodegeneration. *J Neuropathol Exp Neurol* **67**, 377–387 (2008).

146. Hoeijmakers, J. H. J. DNA Damage, Aging, and Cancer. *New England Journal of Medicine* **361**, 1475–1485 (2009).

147. Markowitz, S. D. & Bertagnolli, M. M. Molecular Basis of Colorectal Cancer. *New England Journal of Medicine* **361**, 2449–2460 (2009).

148. Robertson, A. B., Klungland, A., Rognes, T. & Leiros, I. DNA Repair in Mammalian Cells. *Cellular and Molecular Life Sciences 2009 66:6* **66**, 981–993 (2009).

149. Raja, S. & van Houten, B. The Multiple Cellular Roles of SMUG1 in Genome Maintenance and Cancer. *International Journal of Molecular Sciences 2021, Vol. 22, Page 1981* **22**, 1981 (2021).

150. Nilsen, H. *et al.* Excision of deaminated cytosine from the vertebrate genome: role of the SMUG1 uracil–DNA glycosylase. *EMBO J* **20**, 4278–4286 (2001).

151. Wibley, J. E. A., Waters, T. R., Haushalter, K., Verdine, G. L. & Pearl, L. H. Structure and specificity of the vertebrate anti-mutator uracil-DNA glycosylase SMUG1. *Mol Cell* **11**, 1647–1659 (2003).

152. Matsubara, M. *et al.* Mutational analysis of the damage-recognition and catalytic mechanism of human SMUG1 DNA glycosylase. *Nucleic Acids Res* **32**, 5291–5302 (2004).

153. An, Q., Robins, P., Lindahl, T. & Barnes, D. E. 5-Fluorouracil Incorporated into DNA Is Excised by the Smug1 DNA Glycosylase to Reduce Drug Cytotoxicity. *Cancer Res* **67**, 940–945 (2007).

154. Matsumoto, Y. *et al.* Synergistic enhancement of 5-fluorouracil cytotoxicity by deoxyuridine analogs in cancer cells. *Oncoscience* **2**, 272–284 (2015).

155. Masters, L., Eagon, S. & Heying, M. Evaluation of consensus scoring methods for AutoDock Vina, smina and idock. *J Mol Graph Model* **96**, 107532 (2020).

156. Charifson, P. S., Corkery, J. J., Murcko, M. A. & Walters, W. P. Consensus scoring: A method for obtaining improved hit rates from docking databases of three-dimensional structures into proteins. *J Med Chem* **42**, 5100–5109 (1999).

157. Waterhouse, A. *et al.* SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic Acids Res* **46**, W296–W303 (2018).

158. Shuvo, M. H., Gulfam, M. & Bhattacharya, D. DeepRefiner: high-accuracy protein structure refinement by deep network calibration. *Nucleic Acids Res* **49**, W147–W152 (2021).

159. Ropp, P. J. *et al.* GypSUm-DL: An open-source program for preparing small-molecule libraries for structure-based virtual screening. *J Cheminform* **11**, 1–13 (2019).

160. MGLTools. Preprint at https://ccsb.scripps.edu/mgltools/.

161. Trott, O. & Olson, A. J. AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J Comput Chem* NA-NA (2009) doi:10.1002/jcc.21334.

162. Eberhardt, J., Santos-Martins, D., Tillack, A. F. & Forli, S. AutoDock Vina 1.2.0: New Docking Methods, Expanded Force Field, and Python Bindings. *J Chem Inf Model* **61**, 3891–3898 (2021).

163. LigPrep. Preprint at (2021).

164. Maestro. Preprint at https://www.schrodinger.com/products/maestro (2021).

165. Friesner, R. A. *et al.* Glide: A New Approach for Rapid, Accurate Docking and Scoring. 1. Method and Assessment of Docking Accuracy. *J Med Chem* **47**, 1739–1749 (2004).

166. Halgren, T. A. *et al.* Glide: A New Approach for Rapid, Accurate Docking and Scoring. 2. Enrichment Factors in Database Screening. *J Med Chem* **47**, 1750–1759 (2004).

167. Compound Sets - NCI DTP Data - NCI Wiki. https://wiki.nci.nih.gov/display/NCIDTPdata/Compound+Sets.

168. Charifson, P. S., Corkery, J. J., Murcko, M. A. & Walters, W. P. Consensus scoring: A method for obtaining improved hit rates from docking databases of three-dimensional structures into proteins. *J Med Chem* **42**, 5100–5109 (1999).

169. Ericksen, S. S. *et al.* Machine Learning Consensus Scoring Improves Performance Across Targets in Structure-Based Virtual Screening. *J Chem Inf Model* **57**, 1579–1590 (2017).

170. Masters, L., Eagon, S. & Heying, M. Evaluation of consensus scoring methods for AutoDock Vina, smina and idock. *J Mol Graph Model* **96**, 107532 (2020).

171. Zhang, Z. *et al.* Structural Basis of Substrate Specificity in Geobacter metallireducens SMUG1. *ACS Chem Biol* **11**, 1729–1736 (2016).

172. Davies, J. & Davies, D. Origins and Evolution of Antibiotic Resistance. *Microbiology and Molecular Biology Reviews* **74**, 417–433 (2010).

173. Blair, J. M. A., Webber, M. A., Baylay, A. J., Ogbolu, D. O. & Piddock, L. J. V. Molecular mechanisms of antibiotic resistance. *Nat Rev Microbiol* **13**, 42–51 (2015).

174. Zaffiri, L., Gardner, J. & Toledo-Pereyra, L. H. History of antibiotics. from salvarsan to cephalosporins. *Journal of Investigative Surgery* **25**, 67–77 (2012).

175. Antimicrobial resistance. https://www.who.int/news-room/fact-sheets/detail/antimicrobial-resistance.

176. Tooke, C. L. *et al.* β-Lactamases and β-Lactamase Inhibitors in the 21st Century. *J Mol Biol* **431**, 3472–3500 (2019).

177. Datta, N. & Richmond, M. H. The purification and properties of a penicillinase whose synthesis is mediated by an R-factor in Escherichia coli. *Biochem J* **98**, 204–209 (1966).

178. Salverda, M. L. M., de Visser, J. A. G. M. & Barlow, M. Natural evolution of TEM-1 β-lactamase: Experimental reconstruction and clinical relevance. *FEMS Microbiol Rev* **34**, 1015–1036 (2010).

179. Matagne, A., Dubus, A., Galleni, M. & Frère, J. M. The β-lactamase cycle: A tale of selective pressure and bacterial ingenuity. *Nat Prod Rep* **16**, 1–19 (1999).

180. Naas, T. *et al.* Beta-lactamase database (BLDB)–structure and function. *J Enzyme Inhib Med Chem* **32**, 917–919 (2017).

181. Matagne, A., Lamotte-Brasseur, J. & Frère, J. M. Catalytic properties of class A β-lactamases: Efficiency and diversity. *Biochemical Journal* **330**, 581–598 (1998).

182. Horn, J. R. & Shoichet, B. K. Allosteric Inhibition Through Core Disruption. *J Mol Biol* **336**, 1283–1291 (2004).

183. Stec, B., Holtz, K. M., Wojciechowski, C. L. & Kantrowitz, E. R. Structure of the wild-type TEM-1 β-lactamase at 1.55 Å and the mutant enzyme Ser70Ala at 2.1 Å suggest the mode

of noncovalent catalysis for the mutant enzyme. *Acta Crystallogr D Biol Crystallogr* **61**, 1072–1079 (2005).

184. Pettersen, E. F. *et al.* UCSF ChimeraX: Structure visualization for researchers, educators, and developers. *Protein Science* **30**, 70–82 (2021).

185. Williams, C. J. *et al.* MolProbity: More and better reference data for improved all-atom structure validation. *Protein Science* **27**, 293–315 (2018).

186. Wang, J., Wolf, R. M., Caldwell, J. W., Kollman, P. A. & Case, D. A. Development and testing of a general Amber force field. *J Comput Chem* **25**, 1157–1174 (2004).

187. Lu, C. *et al.* OPLS4: Improving Force Field Accuracy on Challenging Regimes of Chemical Space. *J Chem Theory Comput* **17**, 4291–4300 (2021).

188. Sherman, W., Day, T., Jacobson, M. P., Friesner, R. A. & Farid, R. Novel Procedure for Modeling Ligand/Receptor Induced Fit Effects. *J Med Chem* **49**, 534–553 (2006).

189. Sherman, W., Beard, H. S. & Farid, R. Use of an Induced Fit Receptor Structure in Virtual Screening. *Chem Biol Drug Des* **67**, 83–84 (2006).

190. Fusani, L., Palmer, D. S., Somers, D. O. & Wall, I. D. Exploring Ligand Stability in Protein Crystal Structures Using Binding Pose Metadynamics. *J Chem Inf Model* acs.jcim.9b00843 (2020) doi:10.1021/acs.jcim.9b00843.

191. Clark, A. J. *et al.* Prediction of Protein-Ligand Binding Poses via a Combination of Induced Fit Docking and Metadynamics Simulations. *J Chem Theory Comput* **12**, 2990–2998 (2016).

192. Humphrey, W., Dalke, A. & Schulten, K. VMD - Visual Molecular Dynamics. *J Mol Graph* **14**, 33–38 (1996).

193. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **12**, 2825–2830 (2011).

194. Frishman, D. & Argos, P. Knowledge-based protein secondary structure assignment. *Proteins: Structure, Function, and Bioinformatics* **23**, 566–579 (1995).

195. Savard, P. Y. & Gagné, S. M. Backbone dynamics of TEM-1 determined by NMR: Evidence for a highly ordered protein. *Biochemistry* **45**, 11414–11424 (2006).

196. Roccatano, D. *et al.* Dynamical aspects of TEM-1 β-Lactamase probed by molecular dynamics. *J Comput Aided Mol Des* **19**, 329–340 (2005).

197. Bös, F. & Pleiss, J. Multiple molecular dynamics simulations of TEM β-lactamase: Dynamics and water binding of the Ω-loop. *Biophys J* **97**, 2550–2558 (2009).

198. Giampaolo, A. di *et al.* On the structural affinity of macromolecules with different biological properties: Molecular dynamics simulations of a series of TEM-1 mutants. *Biochem Biophys Res Commun* **436**, 666–671 (2013).

199. Meneksedag, D., Dogan, A., Kanlikilicer, P. & Ozkirimli, E. Communication between the active site and the allosteric site in class A beta-lactamases. *Comput Biol Chem* **43**, 1–10 (2013).

200. Huang, L., So, P.-K., Chen, Y. W., Leung, Y.-C. & Yao, Z.-P. Conformational Dynamics of the Helix 10 Region as an Allosteric Site in Class A β-Lactamase Inhibitory Binding. *J Am Chem Soc* **142**, 13756–13767 (2020).

201. Galdadas, I. *et al.* Allosteric communication in class a β-lactamases occurs via cooperative coupling of loop dynamics. *Elife* **10**, (2021).

202. Wienen-Schmidt, B., Oebbeke, M., Ngo, K., Heine, A. & Klebe, G. Two Methods, One Goal: Structural Differences between Cocrystallization and Crystal Soaking to Discover Ligand Binding Poses. *ChemMedChem* **16**, 292–300 (2021).

203. Krishnan, N. P., Nguyen, N. Q., Papp-Wallace, K. M., Bonomo, R. A. & van den Akker, F. Inhibition of Klebsiella β-lactamases (SHV-1 and KPC-2) by avibactam: A structural study. *PLoS One* **10**, 1–13 (2015).

204. Kaynak, B. T., Bahar, I. & Doruker, P. Essential site scanning analysis: A new approach for detecting sites that modulate the dispersion of protein global motions. *Comput Struct Biotechnol J* **18**, 1577–1586 (2020).

205. Kalp, M., Buynak, J. D. & Carey, P. R. Role of E166 in the imine to enamine tautomerization of the clinical β-lactamase inhibitor sulbactam. *Biochemistry* **48**, 10196–10198 (2009).

206. Doucet, N., de Wals, P. Y. & Pelletier, J. N. Site-saturation mutagenesis of Tyr-105 reveals its importance in substrate stabilization and discrimination in TEM-1 β-lactamase. *Journal of Biological Chemistry* **279**, 46295–46303 (2004).

207. Lim, D. *et al.* Crystal structure and kinetic analysis of β-lactamase inhibitor protein-II in complex with TEM-1 β-lactamase. *Nature Structural Biology 2001 8:10* **8**, 848–852 (2001).

208. Zafaralla, G., Manavathu, E. K., Lerner, S. A. & Mobashery, S. Elucidation of the role of arginine-244 in the turnover processes of class A .beta.-lactamases. *Biochemistry* **31**, 3847–3852 (1992).

209. Yang, J., Li, Q. & Bian, L. Spectroscopic analysis and docking simulation on the recognition and binding of TEM-1 β-lactamase with β-lactam antibiotics. *Exp Ther Med* **14**, 3288–3298 (2017).

210. del Sol, A., Fujihashi, H., Amoros, D. & Nussinov, R. Residue centrality, functionally important residues, and active site shape: Analysis of enzyme and non-enzyme families. *Protein Science* **15**, 2120–2128 (2006).

211. Amitai, G. *et al.* Network analysis of protein structures identifies functional residues. *J Mol Biol* **344**, 1135–1146 (2004).

212. Díaz, N., Sordo, T. L., Merz, K. M. & Suárez, D. Insights into the acylation mechanism of class A β-lactamases from molecular dynamics simulations of the TEM-1 enzyme complexed with benzylpenicillin. *J Am Chem Soc* **125**, 672–684 (2003).

213. Medina-Franco, J. L. Grand Challenges of Computer-Aided Drug Design: The Road Ahead. *Frontiers in Drug Discovery* **0**, 2 (2021).

214. Hart, K. M., Ho, C. M. W., Dutta, S., Gross, M. L. & Bowman, G. R. Modelling proteins' hidden conformations to predict antibiotic resistance. *Nature Communications 2016 7:1* **7**, 1–10 (2016).

215. Damry, A. M. & Jackson, C. J. The evolution and engineering of enzyme activity through tuning conformational landscapes. *Protein Engineering, Design and Selection* **34**, 1–6 (2021).

216. Williamson, M. P. Using chemical shift perturbation to characterize ligand binding. *Prog Nucl Magn Reson Spectrosc* **73**, 1–16 (2013).