Title

Development of Computer Aided Drug Design Algorithms and Application to be APOBEC3 Family of Proteins

Permalink

https://escholarship.org/uc/item/9c66d756

Author

Wagner, Jeffrey Robert Rothfeld

Publication Date

2018

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA, SAN DIEGO

Development of Computer Aided Drug Design Algorithms and Application to the APOBEC3 Family of Proteins

A dissertation submitted in partial satisfaction of the requirements for the degree of Doctor of Philosophy

in

Chemistry

by

Jeffrey Robert Rothfeld Wagner

Committee in Charge:
Professor Rommie E. Amaro, Chair
Professor Michael K. Gilson, Co-Chair
Professor Ruben Abagyan
Professor Elizabeth A. Komives
Professor Wei Wang

2018

The Dissertation of Jeffrey Robert Rothfeld Wagner is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

____

____

____

____
Co-Chair

____
Chair

University of California, San Diego

2018

DEDICATION

For my mother and father, who raised us in a warm, loving household and instilled the values that I am proud to carry with me today. My highest goal in life is to return to the world the love and care you showed me.

For Christine, who boldly led the way through life.

For William, my mentor for 17 years.

EPIGRAPH

Don’t you see the plants, the birds, the ants and spiders and bees going about their individual tasks, putting the world in order, as best they can? And you’re not willing to do your job as a human being? Why aren’t you running to do what your nature demands?

Marcus Aurelius, Meditations

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGEMENTS

Graduate school has been an amazing journey. My mentors Rommie Amaro, Michael Gilson, Nagarajan Vaidehi, Abhinandan Jain, and Adam Landsberg have imparted to me lessons that have made me a better scientist and person. It has been my pleasure to learn under their mentorship. I would also like to thank my all my UCSD colleagues, but especially Ozlem Demir, Christopher Lee, Dan Mermelstein, Jesper Sørensen, Robert Swift, Victoria Feher, Robert Malmstrom, and Christopher Churas. Together we have inched forward the boundaries of knowledge and steered through the rocks of academia.

Chapter 2 is a modified reprint of the material as it appears in "Wagner, J.**,** Sørensen, J., Hensley, N., Wong, C., Zhu, C., Perison, T., Amaro, R.E. (2017) POVME3.0: Software for Mapping Binding Pocket Flexibility, *Journal of Chemical Theory and Computation*. doi: 10.1021/acs.jctc.7b00500". The dissertation author was the primary investigator and author of this paper.

Chapter 3 is a writeup of research done with Ozlem Demir and Rommie Amaro, in collaboration with the Harris and Harki labs at the University of Minnesota.

Chapter 4 is a modified, partial reprint of the material as it appears in "Wagner, J.R., Lee, C.T., Durrant, J.D., Malmstrom, R.D., Feher, V.A., Amaro, R.E. (2016) Emerging Computational Methods for the Rational Discovery of Allosteric Drugs, *Chemical Reviews* doi: 10.1021/acs.chemrev.5b00631". The dissertation author was one of two primary authors of the paper.

Chapter 5 is a writeup of research done with Rommie Amaro, Michael Gilson, Shuai Liu, Christopher Churas, and Robert Swift of the Drug Design Data Resource (D3R).

VITA

2008	Rose Hills Undergraduate Research Fellow Landsberg Lab, Claremont Joint Science Department
2009 - 2011	Summer Undergraduate Research Fellow Caltech / NASA-JPL
2011	Bachelor of Arts in Chemistry and Physics Claremont McKenna College
2011 - 2013	Research Associate Vaidehi Lab at City of Hope
2015 - 2016	Graduate Student Intern Pfizer
2016 - 2018	Graduate Student Researcher Drug Design Data Resource
2018	Doctorate in Chemistry and Biochemistry University of California, San Diego

PUBLICATIONS

Shi, K., Demir, .Ö, Carpenter, M. A., **Wagner, J. R.**, Kurahashi, K., Harris, R. S., Amaro, R. E., Aihara, H. (2017) Conformational Switch Regulates the DNA Cytosine Deaminase Activity of Human APOBEC3B, *Nature Scientific Reports* doi: 10.1038/s41598-017-17694-3

Wagner, J. R., Sorensen, J., Hensley, N., Wong, C., Zhu, C., Perison, T., Amaro, R.E. (2017) POVME3.0: Software for Mapping Binding Pocket Flexibility, *Journal of Chemical Theory and Computation*.doi: 10.1021/acs.jctc.7b00500

Wagner, J. R., Lee, C.T., Durrant, J.D., Malmstrom, R.D., Feher, V.A., Amaro, R.E. (2016) Emerging Computational Methods for the Rational Discovery of Allosteric Drugs, *Chemical Reviews* Article ASAP doi: 10.1021/acs.chemrev.5b00631

Yu, D., Sousa, K., Mattern, D., **Wagner, J. R.**, Fu, X., Vaidehi, N., Forman, B.M., Huang, W. (2015) Stereoselective synthesis, biological evaluation, and modeling of novel bile acid-derived G-protein coupled bile acid receptor 1 (GP-BAR1, TGR5) agonists, *Bioorganic & Medicinal Chemistry*, 23(7) 1613-1628

Larsen, A.B., **Wagner, J. R.**, Kandel S., Salomon-Ferrer R., Vaidehi N., Jain A. (2014) GneimoSim: A Modular Internal Coordinates Molecular Dynamics Simulation Package. *Journal of Computational Chemistry*. 35(31):2245-2255. doi:10.1002/jcc.23743.

Larsen, A. B., **Wagner, J. R.**, Jain, A., & Vaidehi, N. (2014). Protein Structure Refinement of CASP Target Proteins Using GNEIMO Torsional Dynamics Method. *Journal of Chemical Information and Modeling*, 54(2), 508-517.

Wagner, J. R., Balaraman, G. S., Niesen, M. J., Larsen, A. B., Jain, A., & Vaidehi, N. (2013). Advanced techniques for constrained internal coordinate molecular dynamics. *Journal of Computational Chemistry*, *34*(11), 904-914.

Gangupomu, V. K., **Wagner, J. R.**, Park, I. H., Jain, A., & Vaidehi, N. (2013). Mapping Conformational Dynamics of Proteins Using Torsional Dynamics Simulations. *Biophysical Journal*, *104*(9), 1999-2008.

Jain, A., Kandel, S., **Wagner, J. R.**, Larsen, A., & Vaidehi, N. (2013). Fixman compensating potential for general branched molecules. *The Journal of Chemical Physics*, *139*(24), 244103.

Park, I. H., Gangupomu, V., **Wagner, J. R.**, Jain, A., & Vaidehi, N. (2012). Structure refinement of protein low resolution models using the GNEIMO constrained dynamics method. *Journal of Physical Chemistry B*, *116*(8), 2365-2375

ABSTRACT OF THE DISSERTATION

Development of Computer Aided Drug Design Algorithms and Application to the APOBEC3 Family of Proteins

by

Jeffrey Robert Rothfeld Wagner

Doctor of Philosophy in Chemistry

University of California, San Diego, 2018

Professor Rommie E. Amaro, Chair

The development of molecular dynamics (MD) simulations builds off the maturing field of structural biology to provide new insight into the mechanisms of disease on an atomic level. However, there are few established methods that use the results of MD to aid the development of novel therapies. This thesis begins by discussing the creation of “POVME 3.0”, a novel method to generate drug design-relevant insights from MD simulations. POVME 3.0 takes as input a MD simulation of a binding pocket of interest, and returns a summary of how the pocket shape changes over time. We then discuss the application of POVME 3.0 and other analysis techniques to the APOBEC3 family of proteins. APOBEC3 proteins are newly discovered drivers of mutation in some tumors, and their inhibition could contribute to cancer treatments. This work studies how physics-based simulations can shed light on the substrate recognition mechanisms of APOBEC3B. Understanding these mechanisms can aid in the discovery of new functional binding sites and modes of therapy. Next, algorithms to detect allostery from protein sequence data are discussed. These algorithms provide a unique data stream that can synergize with physics-based methods to strengthen our understanding of protein function and help develop therapeutic modulators. Finally, we discuss CELPP, a community-driven analysis of computer-aided drug design algorithms, which aims to improve the quality of predictive models in drug design.

Chapter 1 : Computer-Aided Drug Design Algorithms

This dissertation is a documentation of efforts to use computer modeling of proteins and other biological macromolecules to address therapeutically relevant questions. Subsequent chapters will go into detail about the methods used and the results obtained in individual studies. This chapter is intended to help the reader gain an understanding of the context of the work.

While modern algorithms can effectively address problems in some areas of engineering and science, the complexity of biology has hindered the creation of methods to directly address medical questions.[1-6] Fields of science and engineering where computer modeling is effectively applied have a few common characteristics. First, the systems being modeled are often homogenous, for example a solid volume element of constant composition in a finite element model of a machine. Second, the underlying physics are well behaved on the scale of the elements, for example equations of heat transfer. Finally, the number of real-world measurements of the system is at least proportional to the number of parameters in the model, such that the unknown model parameters can be calibrated to match real-world results.

Biological macromolecules differ from these well-defined simulation subjects in that the component pieces (molecules) are highly heterogeneous, the underlying physics (quantum mechanics) are configuration-dependent and expensive to compute accurately, and the number of model parameters is far greater than the available real-world data. Successful use of computation in drug discovery therefore requires clever strategies to work around these shortcomings.

To address the first and second problems, scientists have worked to discretize the elements of a molecular simulation into readily-computable units. In theory, quantum chemical calculations allow scientists to model each atom accurately. In practice, the cost of quantum mechanical computations is prohibitive. A popular strategy is to work from the bottom up, by fitting rapidly-computable molecular energy functions to energies derived from quantum chemistry calculations. This approach led to the development of the molecular “force fields”, which were a fundamental breakthrough that enabled successful use of molecular dynamics simulations. The reasoning behind force fields is that one cannot avoid the variety of atomic behavior in different geometric and charge configurations, but study can yield general patterns that, to a degree, are able to model the behavior of atoms correctly. These general patterns provide the parameters that feed into modern molecular mechanics force fields. The development of these force fields is a great undertaking, and while commonly used parameters such as those for amino acids and solvents have enjoyed heavy investment in their development, small molecule force fields must use a limited number of calculations to cover a large area of chemical space and, as a result, are considerably less accurate.[7-18]

The third problem, that of having more unknown model parameters than real-world measurements, continues to plague computational modeling of biological macromolecules. One example of this discrepancy is the comparison of a physics-based model of a protein to experiments on its biochemical activity. To measure the effect of a change in a protein requires significant material investment. Even when performed, the experiments that measure biochemical activity are sensitive to factors such as temperature, solvent composition, pH, biological cofactors, and other sources of error. Therefore, it is hard to “recreate” the same system in a computer model as in the real world. Because there are so many interacting variables, it is infeasible to measure the effects of changes in all of them, and therefore it is difficult to figure out which model parameters are responsible for resulting inaccuracies. The net effect of this problem is that physics-based protein model parameters are hard to estimate or calibrate, even given a large amount of real-world data.[19, 20] The result is that it is currently not feasible to compute most real-world biochemical values from computer simulations. Strategies to get around this problem include relative comparison of two systems, for example predicting whether a mutation will increase or decrease protein activity rather than the predicting exact numerical difference it would cause, or the reporting of qualitative results from simulations, which experimentalists and domain experts can interpret in the appropriate context.

The author's opinion is that the most efficient way to improve the field of in-silico biochemical simulations is to improve the quality of force fields, develop new analysis techniques for the results of these models that make them more relevant to the real world, and incorporate new data streams to allow for a more favorable ratio of unknown model parameters to real-world data. The projects chosen by the author during graduate study intersect these areas, and the following chapters will discuss them in greater depth.

Chapter 2 of the thesis discusses the development and application of POVME3.0.[21-23] POVME is a program that measures the shape of a protein binding pocket through the course of a molecular dynamics simulation. Version 3.0 is notable as it incorporates high-level methods such as clustering and principal component analysis to the analysis of binding pocket shapes. Clustering is particularly useful in translating the large number of protein conformations that result from a molecular dynamics simulation into a few representative inputs for an ensemble docking procedure. In the context of accuracy, one strength of this approach is that ensemble docking can test its input structures for their ability to reproduce real-world data. In that way, the uncertainty introduced in MD by force fields inaccuracies can be corrected by pruning potentially insignificant binding site conformations from the dataset. This is a highly valuable capability, as there are a few ways to compare simulation with experiment, and this procedure connects the two using a type of data (small molecule binding) that is highly relevant to developing therapeutics.

Chapter 3 of this thesis discusses the application of MD simulations and other forms of structural bioinformatics to the study of APOBEC3B (A3B), a DNA-altering enzyme. A growing body of evidence links A3B to the progression of breast cancer tumors, therefore it is a valuable target for study and potentially a point of application for therapy. This work puts forward a potential structure of the wild-type A3B, which was generated by reverting mutations made by experimentalists to crystallize the protein for X-ray diffraction. MD simulations of this wild-type model were run with and without the substrate DNA, and with a modified DNA strand where the base in the active site is instead RNA. Comparison of these simulations to each other and to previous published biochemical work on the APOBEC family yields potential insights into the mechanism of substrate recognition. In the context of this thesis, this work is notable for showing the complexity of applying computer models to what is, in real life, a highly complex phenomenon (cancer). It also forces a reconsideration of the meaning of "correct" in the context of real-world application. While aspects of the model may be quantitatively incorrect, it is possible that a qualitative phenomenon seen in the simulation could lead to the discovery of a true determinant of protein function. In doing so, a model that is not entirely accurate may increase the odds of a breakthrough when combined with traditional wet-lab experiments.

Chapter 4 is a partial reprint of a review written on methods to discover allostery using computer modeling of proteins.[24] Allostery is a concept in protein dynamics in which an event at one location in the protein structure can affect behavior far away. Allostery has been a rising topic in drug development, as there are many advantages to developing a molecule that bind at a alternate site in a protein structure. The full review discusses a variety of methods that use physical and geometric modeling to make predictions of allosteric sites. However, the work was largely collaborative, and the author felt that it would be most valuable to highlight their own contributions to the manuscript. Therefore, chapter 4 discusses the algorithms that use protein sequence data to predict allosteric sites in proteins, as well as their implementations. This work is particularly interesting in the context of this dissertation as it deviates from physics-based modeling and highlights the potential synergies of using multiple forms of real-world data to make predictions. As stated above, it is difficult to find real-world measurements that can correlate directly to physics-based models, and so the introduction of a data stream based on the sequence and evolution of protein structures adds a unique and valuable angle to therapeutic modeling efforts.

This dissertation concludes with the discussion of the upcoming Continuous Evaluation of Protein Ligand Pose Prediction (CELPP) Challenge. CELPP is a highly ambitious project to benchmark docking algorithms. It fills a critical gap in the field of computer aided drug design by rigorously measuring the accuracy of protein-ligand pose prediction methods. These methods have been traditionally hard to benchmark due to dataset bias, fundamentally different approaches to the problem, and the large number of model parameters that researchers must consider when undertaking prediction efforts. CELPP is a literal application of the framework of modeling described at the beginning of this chapter, in that it aims to enumerate the model parameters involved in ligand pose prediction, and search for optimal performance within the space of possible algorithms and options. As there is no single definition of a docking workflow, CELPP is configured to facilitate community participation, and it is made to integrate seamlessly with servers run by outside groups which generate predictions. Further, the standardization required by such a large-scale benchmark encourages the automation of workflows which previously required human internvention. Beyond providing a constant system to benchmark, removing the human element from the pose prediction process enables the scaling of best-in-class pose prediction workflows to a wider variety of diseases and therapeutic areas.

Chapter 2 : POVME 3.0: Software for Mapping Binding Pocket Flexibility

Jeffrey R. Wagner[1], Jesper Sørensen[1], Nathan Hensley[1], Celia Wong[1], Clare Zhu[1], Taylor Perison[1], and Rommie E. Amaro[1,2]

1 – Department of Chemistry and Biochemistry; University of California, San Diego
2 – National Biomedical Computation Resource; University of California, San Diego

Abstract

We present a substantial update to the open-source POVME binding pocket analysis software. New capabilities of POVME 3.0 include a flexible chemical coloring scheme for feature identification, post-analysis tools for comparing large ensembles of pockets (*e.g.,* from molecular dynamics simulations), and the introduction of scripts and methods that facilitate binding pocket comparison and analysis. We envision the use of this software for visualization of binding pocket dynamics, selection of representative structures for ensemble docking, and incorporation of molecular dynamics results into ligand design efforts.

Introduction

Shape complementarity between a ligand and a binding pocket is a central concept in rational drug design. For this reason, many early structure-guided drug design efforts focused on developing tools to determine which molecules fit into a given binding cavity.[25] While these techniques gained widespread appeal, scientists have since realized that protein-ligand binding is not so much a question of rigid fit as it is question of complementarity between the energetic landscapes of the protein, ligand, and solvent.

Improvements in structure-guided virtual screening have attempted to close the gap between rigid docking methods and flexible thermodynamic reality. One of the more difficult steps in this effort is the handling of molecular flexibility.[26] While algorithms have been developed to efficiently sample a small molecule ligand's conformational landscape,[27] [7] [14] [9] [13] [12] [28] [29] proteins are considerably more complex and proper handling of their flexibility is a more challenging question. [30] [26] [31] [32]

The "ensemble docking" method is used to bridge the gap between the available rigid docking techniques and existing models of protein flexibility. [33] [34] [35] This technique allows researchers to integrate the results of another powerful tool in biophysics research - molecular dynamics (MD) simulations - into drug design work. MD simulations can provide hundreds of thousands of snapshots of a protein's conformation through the course of thermal motion. While it is currently computationally intractable to perform docking on these hundreds of thousands of structures, is it possible to do so for tens of structures. In the ensemble docking method, a large set of protein structures is filtered to a smaller representative set, which is selected to preserve the full range of observed conformational diversity.[36] Performing docking on each member of this smaller representative set is therefore feasible, and should ideally yield the same information as docking to every single snapshot of the protein from the simulation.

The process by which a representative set is selected remains an open question.[35] [34] [37] Inherent in the process of representative selection is the concept of finding meaningful differences between the binding sites in different structures, and ensuring that all of these differences are represented in the reduced set of structures. If done correctly, this categorization of differences should also be useful in itself. Viewing a human-interpretable summary of the major areas and types of variation in a binding pocket would make researchers more efficient and effective in answering a range of scientific questions. For example, a visual summary of binding pocket differences around a promising ligand can inform drug designers about new directions of scaffold functionalization. Further, establishing the characteristics of a target binding pocket that distinguish it from other, similar pockets can enable ligand design with high specificity and fewer off-target effects. As computational methods become more powerful, finding correlations between the binding site shape and distant functional regions of a target protein can enable the design of allosteric ligands.

Previous research has been done on the topic of pocket-studying algorithms.[37] [38] In discussing the context of this work, it is important to draw a distinction between “pocket analysis” and “pocket detection”. POVME is a pocket analysis tool. Pocket analysis is the process of characterizing the shape and flexibility of a cavity in detail. Pocket detection is the process of finding druggable cavities on a protein structure where small molecule ligands might bind. Though these processes seem similar, each is best suited to different mathematical representations. For example, a pocket detection algorithm might rely largely on comparison of a user’s query pocket to a set of known druggable pockets. Such a comparison algorithm would favor easily-rotatable (or even rotation-invariant) representations of binding pockets to accurately perform comparisons independent of reference frame. However, pocket analysis algorithms must preserve fine detail and be able to analyze thousands of frames quickly. The generalizations required to create a rotation-invariant representation of a pocket lead to a significant loss of detail and may incur a high computational cost. Therefore, while compatibility between the tools should be a consideration for creating scientific workflows, one single tool is unlikely to be best for both pocket detection and analysis. Readers interested in pocket representation styles are directed to a separate publication.[39]

The conformational flexibility of a binding pocket can be investigated by studying the differences between many related structures, from sources such as molecular dynamics snapshots, crystallography under different conditions, and homologous protein structures. Previous work has been done to investigate the different approaches that can be used to generate meaningful protein conformations for pocket analysis.[40] However, there is not a single standard for the definition of a region of space as a "binding pocket". Furthermore, it is possible that given the range of reasons for studying a binding pocket and the geometric diversity of cavities where drugs bind, a single standard may not even be appropriate. For example, rules which work for deep pockets may not work well for shallow pockets,[41] and parameters aimed at predicting small molecule druggability may be unsuitable for finding peptide binding sites.

Different pocket analysis programs represent binding pockets in different ways - the two most popular representation types are voxel/grid-based and alpha sphere-based. Although both are suitable for visualizing pocket shape, POVME employs a voxel/grid-based pocket representation, as we have found this to better enable pocket comparison. An in-depth discussion of the advantages of grid-based shape methods is available in the publication of another pocket-analysis tool, TRAPP.[42]

A major disagreement in the field of pocket definition arises from the variety of methods that different programs use to define the boundary of a pocket. While most programs are consistent in how they represent the buried portions of cavities (stopping pocket definition at protein atoms and excluding channels too narrow to host a ligand), existing methods diverge in the logic used to define a boundary at the surface-exposed end of a pocket. This is a long-studied issue in pocket definition, and is sometimes referred to as a "can of worms" problem.[43] Some algorithms terminate the pocket when new points are no longer adjacent to a ligand or pocket-lining residue atom.[42] Others draw numerous vectors out from each possible pocket voxel in different directions and may require a certain fraction of these vectors to intersect a protein atom within a cutoff distance.[44] [45] [46] Another option, employed in POVME[22] and other programs [47] [48] is to "gift wrap" the protein with a convex mesh and exclude all voxels outside the mesh from being defined as part of the pocket.

These differences make it difficult to establish a meaningful definition of "volume" and hinder the rigorous comparison of pocket shapes. The example workflows bundled with POVME 3.0 show best practices for different situations, including the disabling of the convex hull algorithm when performing analysis for quantitative comparison.

Various tools have been developed for the analysis of binding pockets, including fpocket[49] [22] [50], TRAPP[42] [40], PocketAnalyzer(PCA)[46], trj_cavity[45], Epock[51], and Volsite.[44] As the use of integrative modeling and data science continues to grow in biomedical research, it is necessary to develop a tool for analysis of binding pocket shapes that is both suitable for immediate visualization, and also able to interface with more complex data analysis tools. We build off of previous developments in this field to create a package that combines precision in pocket analysis with the ability to integrate results into larger data workflows.

In this paper, we present POcket Volume MEasurer (POVME)[21] 3.0 as a tool for analysis of flexible protein cavities. Version 3.0 contains many additional capabilities, including: post-processing tools to perform clustering and principal component analysis; a chemical coloring scheme for defining pocket features; python classes for custom analyses of pocket shape output; pre-built workflows for a variety of tasks; and easy installation using the pypi package index.

Methods

All of the functionality available in POVME2.0 has been maintained, and readers interested in a detailed description are directed to that paper.[22] The new features in POVME3.0 are detailed in this section.

Ligand-based pocket definition

POVME relies on a user-defined inclusion region to define the boundary of the pocket of interest, similar to the Maximum Encompassing Region in Epock.[51] Based on feedback for POVME 2.0, we learned that users frequently found the existing region-definition methods to be unwieldy. For this reason, we added three new features for pocket definition, including defining the pocket based on the 1) residue name of a ligand present in the trajectory, 2) a saved POVME shape file (a 3xN numpy array of grid points), and 3) 3D cylinders (in addition to the previously implemented boxes and spheres). The ligand-based pocket definition is likely to be the most popular option, especially for defining appropriate inclusion regions to analyze tight pockets. When given the DefinePocketByLigand keyword and a ligand residue name, POVME will pre-process the trajectory, map each ligand atom in each frame to its nearest grid point, and define those grid points as the seed region for all frames. It will then grow this seed region 3 Angstroms out in each direction, and define that superset of points as the inclusion region.

Convex hull options

The ConvexHullExclusion keyword can be set to 4 options: “each”, “none”, “first”, or “max”. The first two options were available in an earlier version of POVME. “none” will forego the convex hull exclusion process altogether, as was standard behavior for the original POVME. “each” will calculate a convex hull for each frame of a trajectory, as was standard behavior for POVME 2.0. It is worth noting that the “each” setting is not advised if the final goal of POVME analysis will include quantitative analysis such as clustering or PCA, as the outer boundary of the pocket may shift each frame, and the magnitude of this shift can dwarf motions inside of the pocket. The other two options are new additions and apply the same convex hull to each frame. “first” applies the convex hull from the first frame in the trajectory, and “max” applies a convex hull drawn around all frames in the trajectory superimposed on each other.

Coloring

A chemical coloring scheme has been implemented to characterize portions of the pocket that can host favorable interactions with small molecule ligands. The coloring scheme is based on the BINANA binding site description algorithm,[52] but has not been validated for a quantitative purpose as it is applied in POVME and so is primarily suited for visualization. Currently, this coloring scheme depicts hydrogen bond donors, hydrogen bond acceptors, aromatic stacking, hydrophobicity, and hydrophilicity. The colors are output as separate POVME maps with variable intensity assigned to each grid point. POVME provides time-averaged color maps after analysis of an entire trajectory.

Because the colors are defined by continuous functions but are only defined at discrete points, the total contribution of each feature (for example, a single O-H donor group, or a single aromatic ring) may vary depending on how the intensity of the 3D function falls on the fixed cartesian grid. For this reason, the total contribution of each feature to the grid is normalized, so that the summed value of the feature's contributions to the color grid is equal to 1. Further, in order to ensure that buried points are not assigned color magnitudes, only points that are defined as part of the pocket in a frame (or are within a skin distance of the surface) will receive these color values.

The magnitude of the hydrogen bond donor color is defined as a gaussian in spherical coordinates, with a center beyond the hydrogen atom as measured along O-H or N-H axis. The magnitude of the hydrogen bond acceptor color is defined as a gaussian in Cartesian coordinates, emanating from the center of all O atoms. The aromatic color is defined in a cylinder above and below aromatic rings, with uniform magnitude along the radius (dropping to 0 at an outer radial cutoff) and with magnitude defined by a gaussian along the height of the cylinder. The magnitude of the hydrophobic color is defined as a gaussian around all C atoms, and the hydrophilic is a gaussian around each N, O, and S.

The pocket coloring scheme is extensible to python programmers, and the POVME package contains the "featureMap" class which enables coloring based on a number of shapes at user-defined atom motifs.

Adjacency and surface

Two boundary-defining colors are also defined as “adjacency” and “surface”. Adjacency represents a thick layer of binding pocket volume near the surface of the protein, and may be of interest in measuring buriedness of voxels. Surface represents a thin layer of volume on top of the protein surface lining the binding pocket, and is used in surface area calculation.

File conventions

POVME3.0 requires a pre-aligned trajectory in PDB format. POVME 3.0 outputs 3 file types: pdb, dx, and npy. The first two are chosen for ease of visualization. Boolean grid data, for example individual pocket volumes and regions where color maps exceed a threshold magnitude, are output in Protein Data Bank “pdb” format. Non-boolean grid data are output in Data Explorer “dx” format, which is compatible with a variety of visualization programs including VMD and PyMol. Examples of such data include the average pocket shape of many frames, or color maps in full detail. Every single-frame .pdb and .dx file output from POVME 3.0 also has a .npy equivalent. The NumPy file format was selected for its efficiency, interconvertibility with other file types, and compatibility with major data analysis packages. Efforts to involve POVME in more elaborate integrative modeling efforts should work directly with these npy files.

Clustering

POVME 3.0 offers scripts to perform pocket shape-based clustering and examples to exhibit their use. Pocket shape clustering is handled in two steps. First, a pairwise binding pocket similarity matrix is generated for all binding pocket structures in the ensemble. Second, this similarity matrix is clustered and useful depictions of the clusters and their differences are created.

The similarity matrix is calculated using the Tanimoto overlap score of each pair of pockets. As the grid points in POVME are defined in the same Cartesian reference frame for each pocket, the Tanimoto score is calculated by counting how many pocket points each pair of pockets has in common, divided by the number of points in either. Therefore, the Tanimoto score of a pair of frames can be at maximum 1 (the two pockets are identical) or at minimum 0 (the two pockets share no volume in common). Alternatively, the similarity matrix can be calculated using the Tversky similarity metric, in which the overlap term is the same but the denominator is the volume of one frame instead of the union of both.

In the clustering step, users may select to use SciPy's hierarchical or k-means libraries.[53] [54] By default, hierarchical clustering is performed, based on SciPy's average linkage implementation. The desired number of clusters can be input manually, otherwise cluster.py will compute the Kelley penalty[55] to determine a reasonable number. For each cluster, cluster.py extracts the representative structure (the pdb structure corresponding to the cluster member with the maximum summed overlap score with all of the others) and generates two dx files depicting 1) the cluster's average pocket shape and 2) the difference between this cluster's average and the entire ensemble's average. VMD scripts are produced to load these volume maps and overlay them on the representative structure for each cluster. Figures are also created to show cluster membership as a function of frame number and to create a "kinetic network" diagram of the clusters, linked by the number of transitions the ensemble took between them.

PCA

Principal component analysis (PCA) is a common tool in data science that has recently been applied to pocket shape analysis.[46] [42] PCA of pocket shapes can serve a variety of purposes. First, it can act as a way to define meaningful subpockets. As subpockets can come in a variety of shapes and sizes, it is difficult to select a single method or heuristic to define them. However, it is possible to find mutually correlated groups of voxels that join or leave the pocket together. These mutually correlated groups are often physically contiguous and represent entire subpockets available for ligands. Second, it is possible to find multiple subpockets present in the same eigenvector, with coefficients that indicate positive or negative correlation with one another. Information such as negatively correlated subpockets may be valuable in ligand design, as it would indicate two areas of the binding pocket that are unfavorable for a ligand to occupy simultaneously. Third, PCA allows researchers to define meaningful axes by which structures can be compared quantitatively, which may be useful in selecting structures and rationalizing differences between families of structures.

PCA in POVME is performed by constructing a matrix of pocket points M, in which rows correspond to the different structures in the ensemble, and columns correspond to individual grid points. For each position i,j in the matrix, M(i,j)=1 if grid point j (for example (10,-7,5)) is defined as part of the pocket in structure i of the ensemble. Otherwise M(i,j)=0. Mean normalization, but not feature scaling, is performed on the columns of this matrix. After eigenvalue decomposition of this matrix, each eigenvector is mapped back to a density map defined at each point on the grid and saved as a dx file. These dx files can be visualized by a number of programs, and the workflow outputs a VMD[56] script to load them all simultaneously and prepare a default visualization. This default visualization loads each eigenvector as a different object and displays regions in green and red to denote positive and negative coefficients respectively.

Common workflows

The POVME 3.0 download contains example workflows that users can adapt to their own data, including combined multiple-trajectory analysis, clustering, and principal component analysis.

Pypi distribution

POVME is now available on the Python Package Index (https://pypi.python.org/pypi/povme). This improvement streamlines the installation and updating process. The POVME source code is also now version controlled on GitHub (https://github.com/POVME/POVME), which makes it easier to download, modify, and manage bug reports and feature requests.

HSP90 MD simulations

Twenty 250-ns molecular dynamics simulations were run beginning from different HSP90 crystal structures in the Protein Data Bank (PDB).[57] These PDB codes were selected on the basis of ligand diversity, structure resolution, and pocket characteristics. The final 20 PDB codes selected are 1BYQ,[58] 1UYF,[59] 1UYI,[59] 1UYL,[59] 2VCI,[60] 2WI7,[61] 2XHR,[62] 2YEJ,[63] 3B26,[64] 3D0B,[65] 3HEK,[66] 3K98,[67] 3K99,[67] 3RKZ,[68] 3RLR,[69] 4CWN,[70] 4FCR,[71] 4LWE,[72] 4R3M,[73] 4W7T.[74] Active site ligands were parameterized using GAFF,[15] with charges derived using Gaussian[75] and the RED server.[76] [77] [78] All crystal waters and ligand counterions were preserved. Sodium ions were added to balance the system charge. Schrodinger protein preparation was used to model missing loops, replace unresolved sidechains, and assign protons at pH 7. The full commandline instruction passed to Schrodinger's prepwizard is provided in the SI. The twenty systems were prepared for simulation using LEaP from the AMBERTOOLS package.[27] The FF99SB force field was used for simulation.[10] [27]The tleap solvatebox command was used to add a TIP4P water box with 10A padding.

The AMBER MD input scripts are provided in the Supporting Information.

Results and Discussion

Coloring scheme

The coloring process is performed by default when POVME 3.0 is run. Figure 2.1 shows two examples of the coloring process. As no appropriate weighting scheme has been determined, the clustering and PCA workflows do not consider the color data (only the pocket shape). However, due to their qualitative utility, the color files are provided as pdb, dx, and npy files for visualization and custom user analysis.

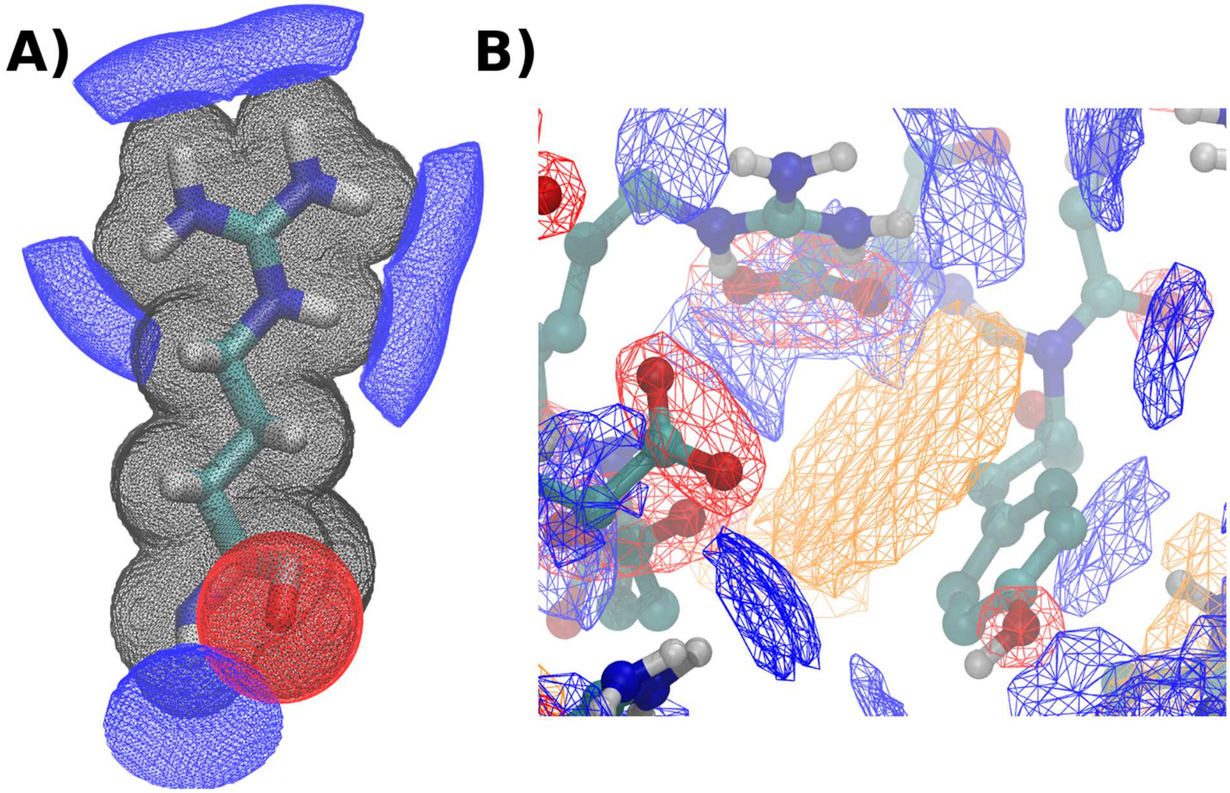


Figure 2.1 : POVME coloring scheme.

Black mesh shows occupancy, blue shows hydrogen bond donor regions, red shows hydrogen bond acceptor, and orange shows pi-stacking. A) The POVME coloring scheme applied at 0.1 Angstrom grid resolution to a single arginine residue from a protein structure. B) The POVME coloring scheme applied to a binding pocket at 0.75 Angstrom resolution, with occupancy not shown.

Validation of pocket similarity metric

We anticipate that researchers will use POVME to guide ligand design based on pocket geometry. Therefore, one of our major scientific objectives is to ensure that the similarity score that POVME reports when comparing binding pockets is related to the similarity of the ligands which fit in those pockets. In other words, if POVME analysis indicates that two pockets are similar, they should bind similar ligands. Conversely, if POVME determines that two pockets are dissimilar, they should bind dissimilar ligands. Establishing such a correlation would provide evidence that POVME's selection of "diverse" pockets from an ensemble of protein structures will enable discovery of diverse ligands.

As a simple study of POVME's pocket similarity metric (Tanimoto scoring), we attempt to use it to distinguish between the same protein crystallized and simulated with 20 different ligands. Each simulation was run for 250 ns, and frames were extracted every 1 ns. To determine the similarities between pockets, POVME was run on each trajectory, and the results were used to make three 20x20 similarity matrices. These matrices show the POVME similarity of the pockets from the first frame of each simulation (Figure 2.2A), the POVME similarity of the pockets from the last frame of each simulation (Figure 2.2B), and the average POVME similarity of all 250 frames taken from each simulation (Figure 2.2C). In order to compare pocket similarity to ligand similarity, it is necessary to compute a ligand similarity matrix. RDKit FingerprintMol objects were generated for each ligand, and the default RDKit similarity metric (Tanimoto) was used to compute a 20x20 ligand similarity matrix (Figure 2.2d).

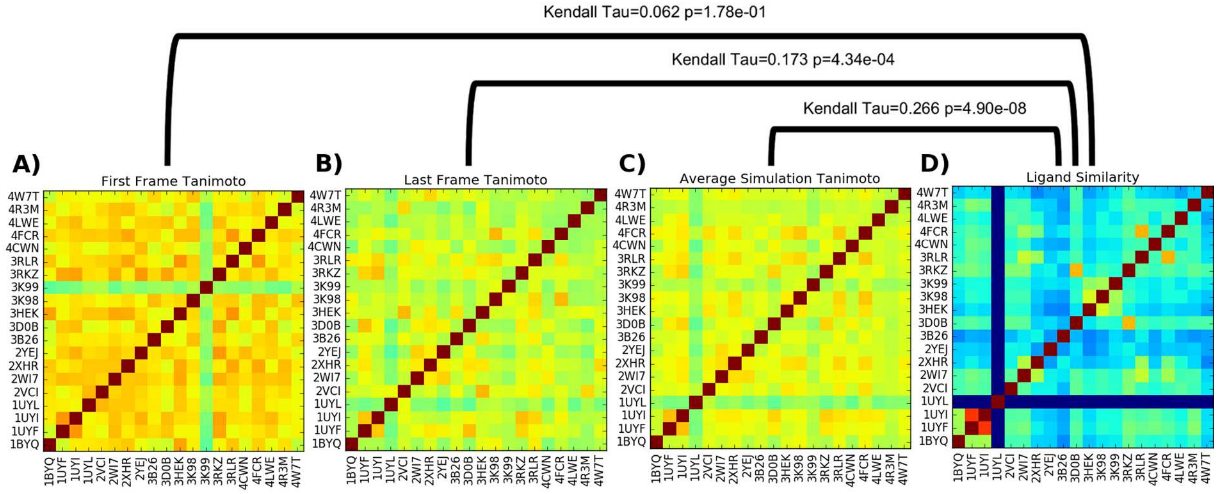


Figure 2.2: Tanimoto similarity matrices

Tanimoto similarity matrices of A) initial simulation pocket shapes (following equilibration), B) final simulation pocket shapes (after 250 ns MD), C) average similarity of all pocket shapes throughout simulation, and D) RDKit chemical fingerprint Tanimoto score. Kendall Tau analyses of the matrices are shown above the brackets.

To compare the information contained in each similarity matrix, the Kendall rank correlation coefficient[79] is employed, which indicates the similarity between two sets of ranked objects. In this case, the ranked objects are pairs of nonidentical HSP90-ligand systems, each denoted by a pair of PDB codes, and they are ranked by their Tanimoto scores. For example, in the ligand similarity matrix (Figure 2.2D), the bright red (1UYI,1UYF) hotspot has the highest nonidentical similarity value (See figure 2.5 for ligand structures). The (1UYI,1UYF) pair therefore has rank 1. The rest of the PDB code pairs are ranked in order of decreasing ligand similarity to create the ordered ligand similarity list. The 1UYL simulation does not have a ligand, therefore it has a similarity score of 0 to all other ligands.

This process is repeated on each pocket similarity matrix to generate the three ordered pocket similarity lists. The Kendall rank correlation coefficient indicates how similar each pair of orderings is, with a maximum possible value of 1 (indicating identical ordering) and a minimum value of -1 (indicating completely opposite ordering). A Kendall Tau of 0 indicates random ordering. Comparing the average simulation Tanimoto similarity matrix (Figure 2.2C) to the ligand similarity matrix (Figure 2.2B) yields a Kendall Tau value of 0.266, with a p of 4.90 x 10^{-8}, indicating moderate agreement with high confidence. Comparisons of only the first and last frame indicate weaker agreement between the rankings; analysis of the last frames of each simulation yields a Kendall Tau of only 0.173, and analysis of the first frames yields a Kendall Tau of 0.062.

The correlation between pocket shape similarity and ligand similarity suggests that using POVME to pick diverse pocket shapes will enable discovery of diverse ligands. To efficiently pick diverse structures, clustering analysis is performed on the complete 5000 x 5000 Tanimoto similarity matrix.

Clustering analysis

The clustering workflow was run on frames taken from the HSP90 trajectories at 1 ns intervals, for a total of 5,000 structures (250 snapshots per simulation x 20 simulations). The workflow is capable of choosing a number of clusters automatically using the Kelley penalty method.[55] However, this number is somewhat arbitrary, and in a project-driven analysis the number of clusters would be better determined by the computational resources available for ensemble docking. As an example, this study sets it to return 15 clusters for ease of visualization, to show how 20 simulations can be reduced. The 15 clusters are numbered 0 to 14, in order of decreasing size. These clusters represent frequently visited pocket shapes (Figure 2.3A). Each frame that is analyzed is assigned to a single cluster. Scientists using POVME to select diverse structures for ensemble docking will be primarily interested in the frames identified as cluster representatives by this step.

As ligand kinetics are increasingly recognized to play an important role in drug efficacy (exemplified by recent interest in slow-k_{off} ligands)[80], understanding the kinetics of ligand-binding pockets also becomes a valuable topic of study. POVME clustering offers a way to discretize pocket conformations, and scientists can study pocket kinetics by observing how the systems transition between clusters. While the clustering process analyzes the trajectories together (as one large concatenated trajectory), the results of clustering can be mapped back over the different systems, and the time evolution of the simulations through the clusters can be studied (Figure 2.3B and C).

In the HSP90 data, we observe that the low-numbered (and therefore larger) clusters contain frames from multiple simulations, while clusters numbered 10 and above are all populated by a few outlier frames from individual simulations (complete data in Figure 2.6). Further, it is observed that the simulations of HSP90 bound to highly similar ligands, 1UYI and 1UYF, are the two largest occupants of cluster 1, but that only 1UYF makes excursions to cluster 9, which exhibits the opening of a side channel. An apo crystal structure uploaded as part of the same publication, 1UYL, starts in the most populated cluster, 0, but quickly transitions to cluster 5, which features a collapsed binding region and is populated exclusively by the apo simulation. The ligand from the 4R3M crystal structure, while sharing limited structural similarity with the 1UYF and 1UYI ligands, populates in small parts clusters 1 and 2, but is found the majority of the time in conformations bordering cluster 6. This cluster represents a higher-volume binding pocket with a unique deep subpocket open (Figure 2.3A). While this paper does not go in-depth on the SAR linking ligand chemotype to pocket conformation, it demonstrates that POVME enables the analysis of ligand-induced changes in protein dynamics.

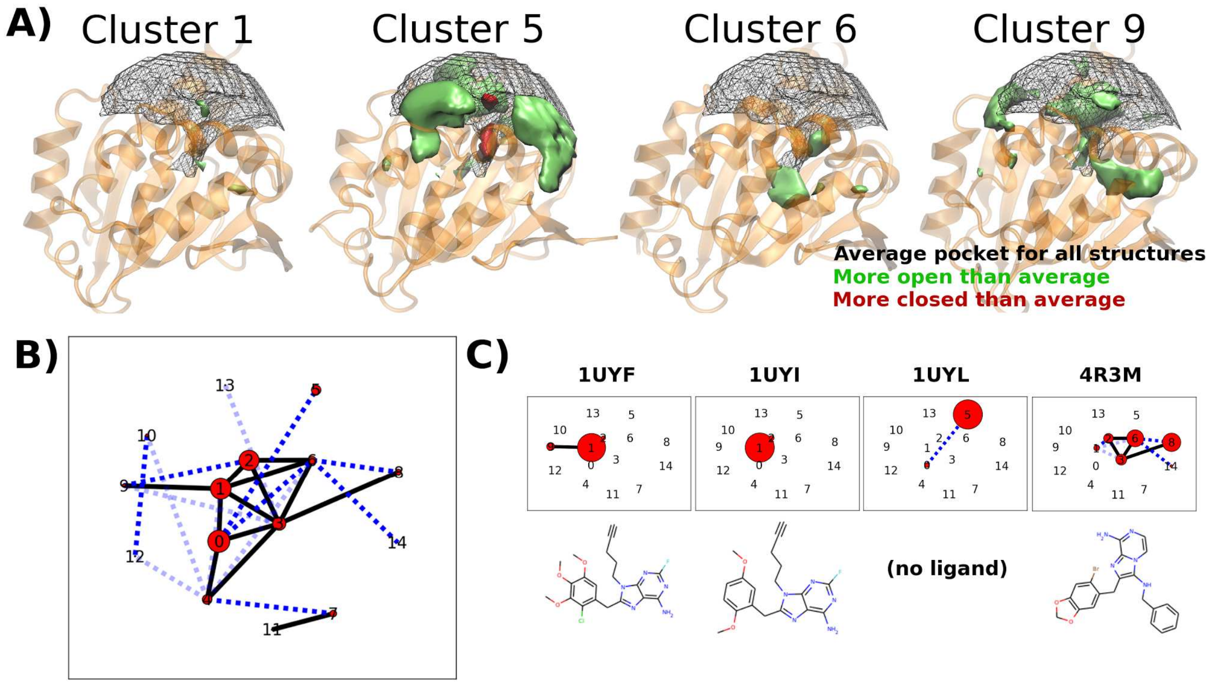


Figure 2.3 : POVME clustering of 20 HSP90-ligand complexes using cluster.py

A) 3D depictions of selected clusters. The representative protein structure for each cluster is shown as a transparent orange cartoon. The average shape for all analyzed frames is shown as a black mesh, with solid shapes (green and red) showing how each cluster is more open or closed than this average. B) "Kinetic network" depiction of the combined dynamics of the 20 HSP90 trajectories. Black numbers indicate cluster index (0-14). Red circles indicate the number of frames assigned to each cluster. Edges indicate the number of transitions observed between clusters in the MD trajectories (light blue dashes = 1 or 2, dark blue dashes = 3 to 5, solid black line = greater than 5). Clusters are arranged in 2D according to a force-based layout, in which each pair of clusters is pulled together by a force proportional to the number of observed transitions. C) Kinetic network depiction of individual trajectories. These network diagrams have the same 2D cluster locations, but only show cluster populations and transitions from individual trajectories. The corresponding ligand is shown below each diagram. 1UYL is the only apo structure, and is the only simulation that visits cluster 5.

Principal Component Analysis

Principal Component Analysis was performed on the 20 HSP90 trajectories. Figure 2.7 shows that the pocket dynamics are complex with regard to subpockets - the first 10 principal components describe only about 30% of the pocket dynamics. However, reviewing the most significant principal components can be informative, as they explain major areas of pocket variation and how they relate to ligand structure. PC1 shows a change in pocket shape corresponding to the interruption of a binding site-adjacent helix (Figure 4A). This change opens a subpocket below the helix. PC2 corresponds to a complete loss of the same helix, and the inward bulging of secondary structure on the far side of the pocket (Figure 2.4B).

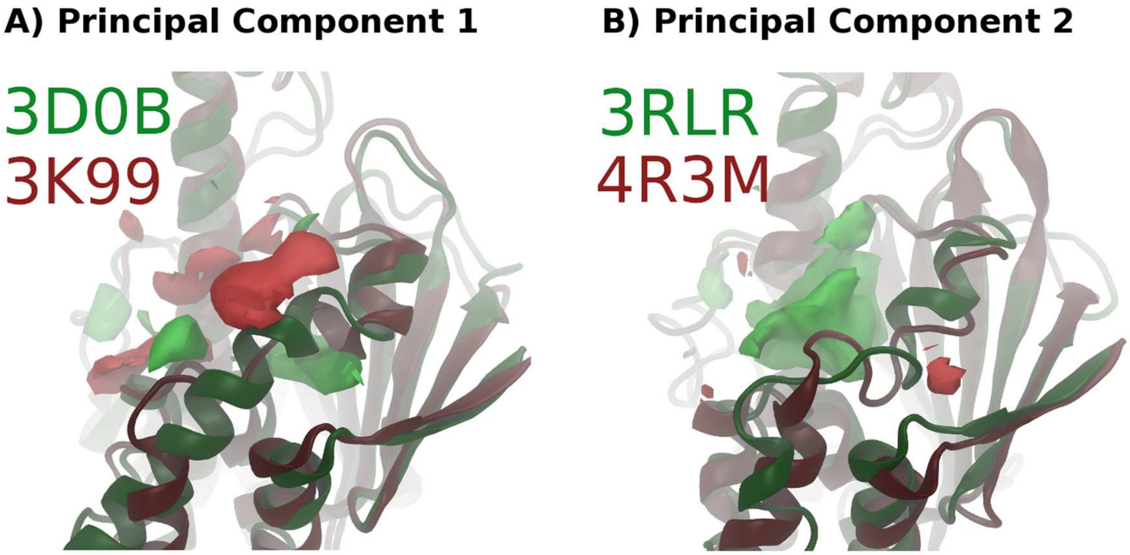


Figure 2.4 : POVME principal component analysis of 20 HSP90-ligand simulations.

PCs were derived from analysis of 250 ns of MD of each system. Bright green and red regions are spatial depictions of each principal component in pocket space - A structure with a high PC value would have the green areas included in the pocket and the red areas closed off. Protein structures are snapshots from simulations corresponding to the high and low values in each PC. A) PC1, showing the final frame of the highest (3D0B, green) and lowest (3K99, red) valued systems. B) PC2, showing the final frame of the highest (3RLR, green) and lowest (4R3M, red) valued systems.

Conclusions

We present POVME 3.0, a substantial update to the POVME package that performs pocket selection for ensemble docking and provides outputs suitable for quantitative analysis. A number of new features have been added, including a chemical coloring scheme for binding pockets, the option to define pocket regions based on the position of a ligand molecule, and detailed manual pocket definition options. Further, post-processing workflows have been provided to perform the principal component and clustering analysis shown in this paper. Finally, POVME 3.0 has been redesigned for distribution on PyPI, simplifying its installation and use.

Great strides in molecular modeling are currently being made, thanks largely to the continued development of open-source software and the standardization of data formats. POVME 3.0 aims to make the field of drug design more open to machine learning techniques by providing a tool that connects MD simulations, pocket shapes, and ligand binding. The workflows for pocket clustering and PCA are initial examples of how POVME 3.0 can interface with statistical learning methods. Pocket shape data will become more valuable when it is combined with other forms of information to, for example, study allostery and correlate pocket shape to ligand structure.

Supplemental Information

User Notes and Best Practices

Effect of structure alignment on pocket analysis

In the course of performing this work, we determined that robust alignment of the protein pocket is a prerequisite for successful POVME analysis. Many trajectory-handling programs such as VMD can perform RMSD alignments, and most default to alignments based on the entire protein (*e.g.,* all alpha carbons). However, some proteins undergo significant domain motion, so care should be taken to perform the alignment such that the binding pocket remains in the same location in Cartesian space. This may require performing an alignment of only the domain containing the binding pocket, or restricting the alignment to a set of pocket-lining residues. During the development and testing of POVME 3.0, inappropriate alignment of the protein trajectory/ensemble was a common problem. Misalignment is usually noticed during clustering and PCA, and is represented by difference regions that line surfaces on opposite sides of the binding pocket. In these cases, one entire face of the pocket is seen to lose volume over its surface, and the opposite face is seen to gain it. This type of change is likely an artifact, adding noise to the interpretation of pocket dynamics. As a solution, we investigated the possibility of providing tools for alignment of pocket shapes, but POVME's voxel-based representation was found to be poorly suited for this task.

We advise against interpretation of scalar volume values

The value that POVME provides for pocket volume is simply the sum of the volumes of the voxels comprising the pocket. Because heuristics are used to define the outer boundary of the pocket, the numerical value of the pocket volume is difficult to meaningfully compare between programs, or even significantly different pockets analyzed by the same program. Users should take caution when comparing the POVME-provided volumes to anything except highly similar pockets. Without knowing how users plan to interpret or compare volume numbers we cannot ensure that they are fit for a specific purpose. Instead, we encourage users to compare pockets in 3D. POVME provides directly visualizable outputs as well as Python functions for performing mathematical operations on sets of pocket shapes. Given the frame-by-frame output files provided by POVME, users with Python knowledge can load the sets of pockets as lists of points, then use POVME functions to compute their difference and output it as a pdb or dx file for visualization.

Inclusion and seed regions must be identically defined for successful post-analysis As the clustering and PCA processes consider variation in pocket shape, it is important that the volume eligible to be part of the pocket is consistently defined for all frames being studied. In other words, post-processing requires that the inclusion and seed regions be identically defined for all of the trajectories being analyzed. The provided workflows take care of this step automatically, by taking as input a user-defined inclusion and seed region. However, when running POVME analysis separately on multiple trajectories with the intent of combining their results in post-analysis, it is essential that their inclusion and seed regions be the same.

Post-processing analysis will not work if the pockets being analyzed have different boundaries. The outer boundary of the pocket is defined both by the edge of the inclusion region, and if the "ConvexHullExclusion" keyword is used, by the convex hull of the protein. When comparing pocket volumes within the same trajectory, users should ensure that the boundary of the pocket is consistently defined. Recalculating a different convex hull for each frame of a trajectory adds noise to quantitative analysis, as the convex hull definition is sensitive to movements of surface residues. Because many pockets widen as they approach the surface of the protein, small changes in how the outer boundary is defined can lead to large numbers of points being added to or removed from the pocket. During quantitative analysis, this large number of variable points will outweigh the smaller changes corresponding to pocket dynamics and shape change inside the cavity.

To instruct POVME to use a single definition of this outer boundary, users should ensure that the ConvexHullExclusion option is set to a keyword other than "each". The default keyword, "none", is recommended. While this choice may lead to a large number of points being defined outside of the pocket, POVME's clustering and PCA scripts focus on *differences* in pocket shape, thus points that lie outside of the protein and are never removed from the pocket do not affect the results of the analysis.

On the inner barrier of a pocket, users should be mindful of another potential source of noise. When a pocket of interest is near another cavity, the protein atoms will sometimes rearrange during MD to join the two. When this joining occurs, the pocket region defined by POVME can become much larger, thereby adding noise to post-processing. Two options to avoid this situation are: 1) If a ligand is present, use the “DefinePocketByLigand” keyword to define the pocket as the area immediately around the ligand, or 2) carefully define inclusion and seed regions so that the unwanted cavity is not included in the analysis.

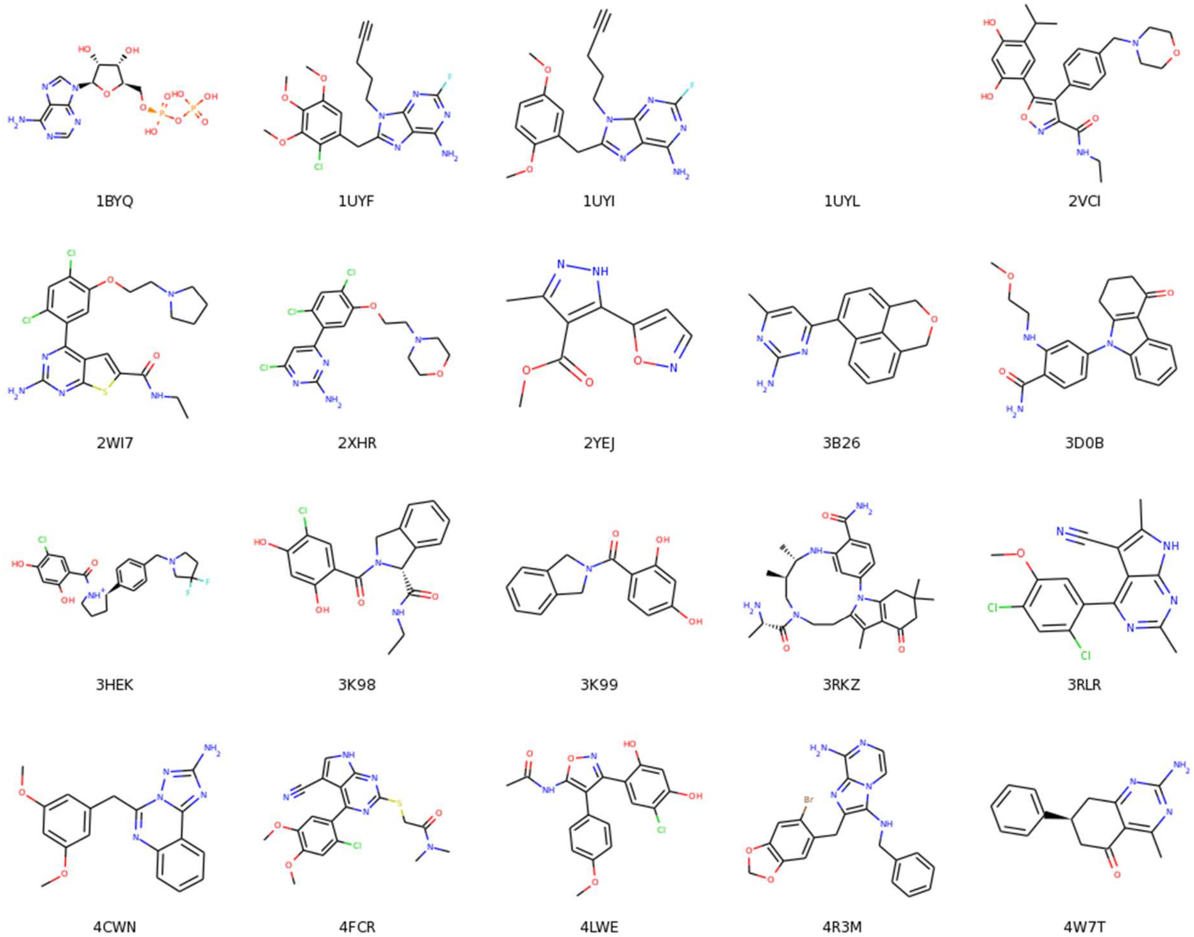


Figure 2.5 : All HSP90 ligands simulated in this work

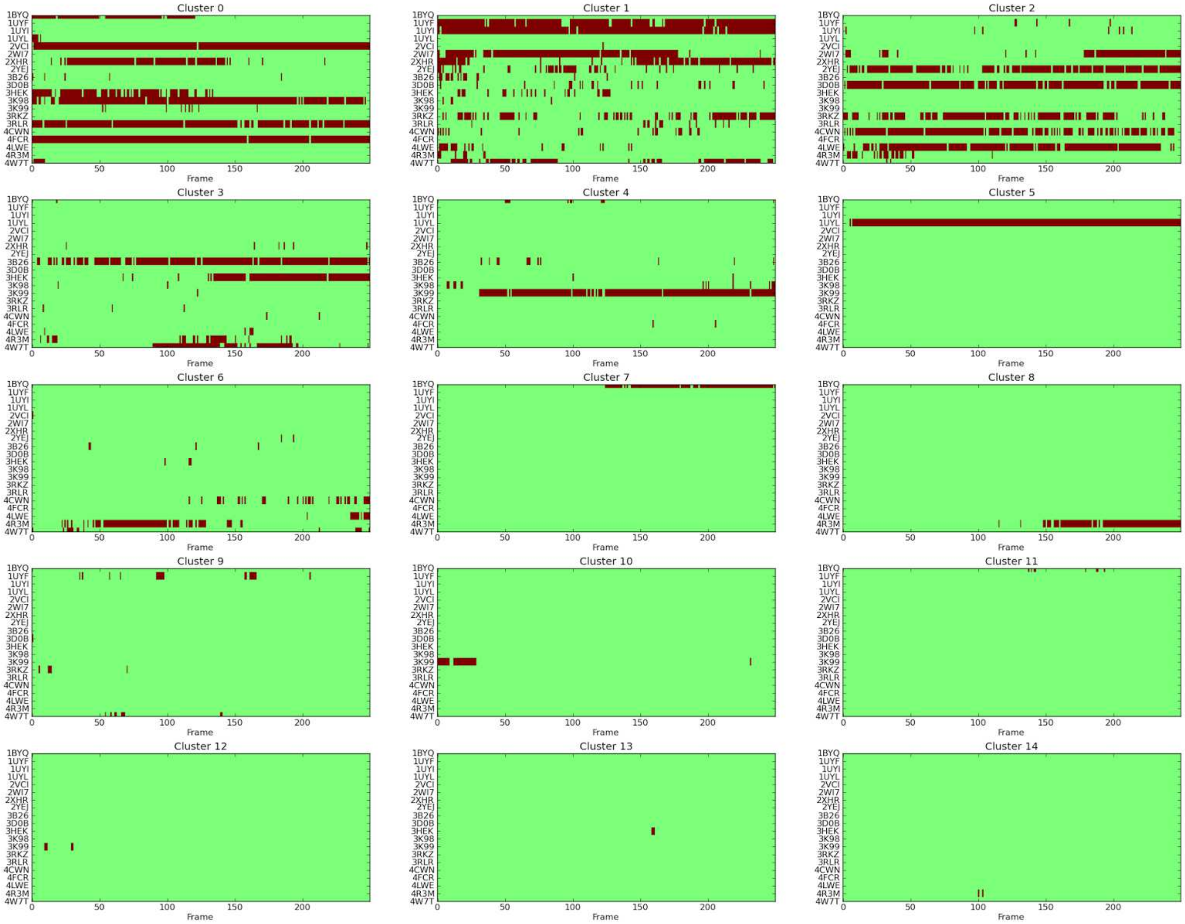


Figure 2.6 : POVME cluster assignments for all frames from HSP90 trajectories.

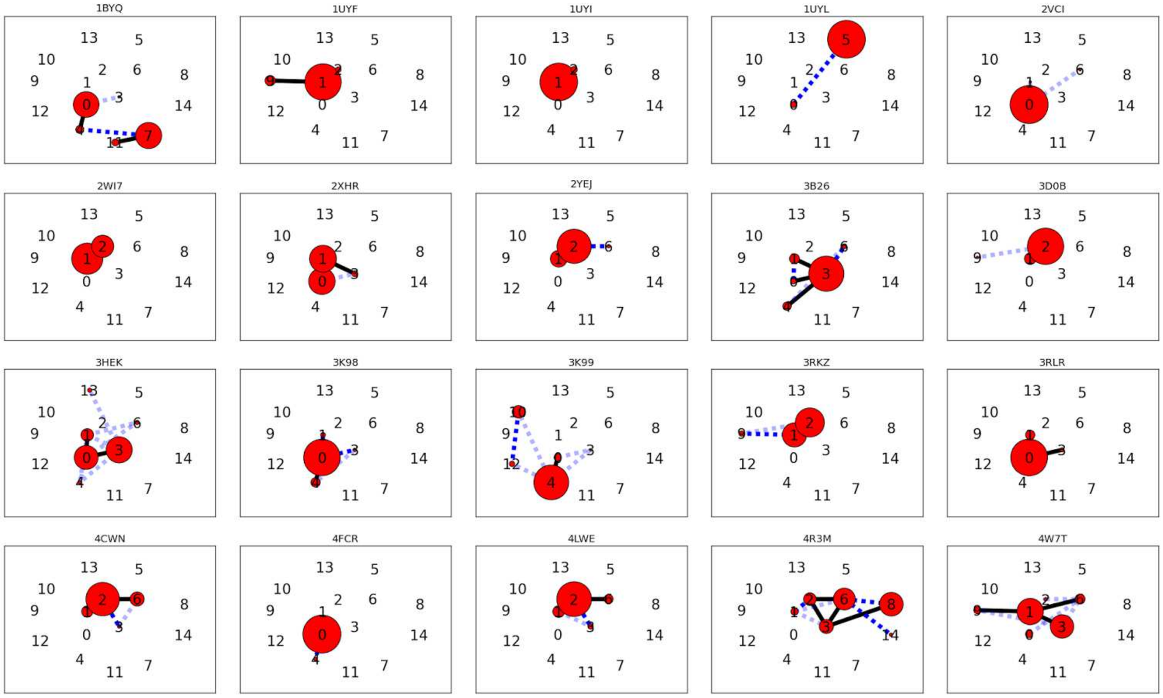


Figure 2.7 : “Kinetic network” depiction of the 20 HSP90 trajectories.

Black numbers indicate cluster index (0-14). Red circles indicate the number of frames assigned to each cluster. Edges indicate the number of transitions observed between clusters in the MD trajectories (light blue dashes = 1 or 2, dark blue dashes = 3 to 5, solid black line = greater than 5). Clusters are arranged in 2D according to a force-based layout, in which each pair of clusters is pulled together by a force proportional to the number of observed transitions.

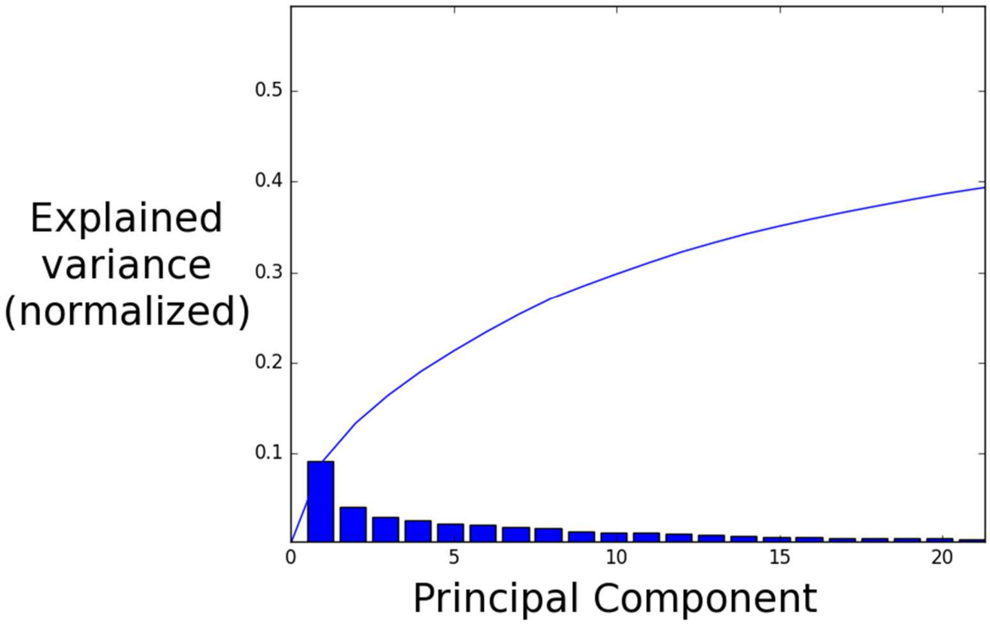


Figure 2.8 : Explained variance plot of Principal Component Analysis of HSP90 trajectories.

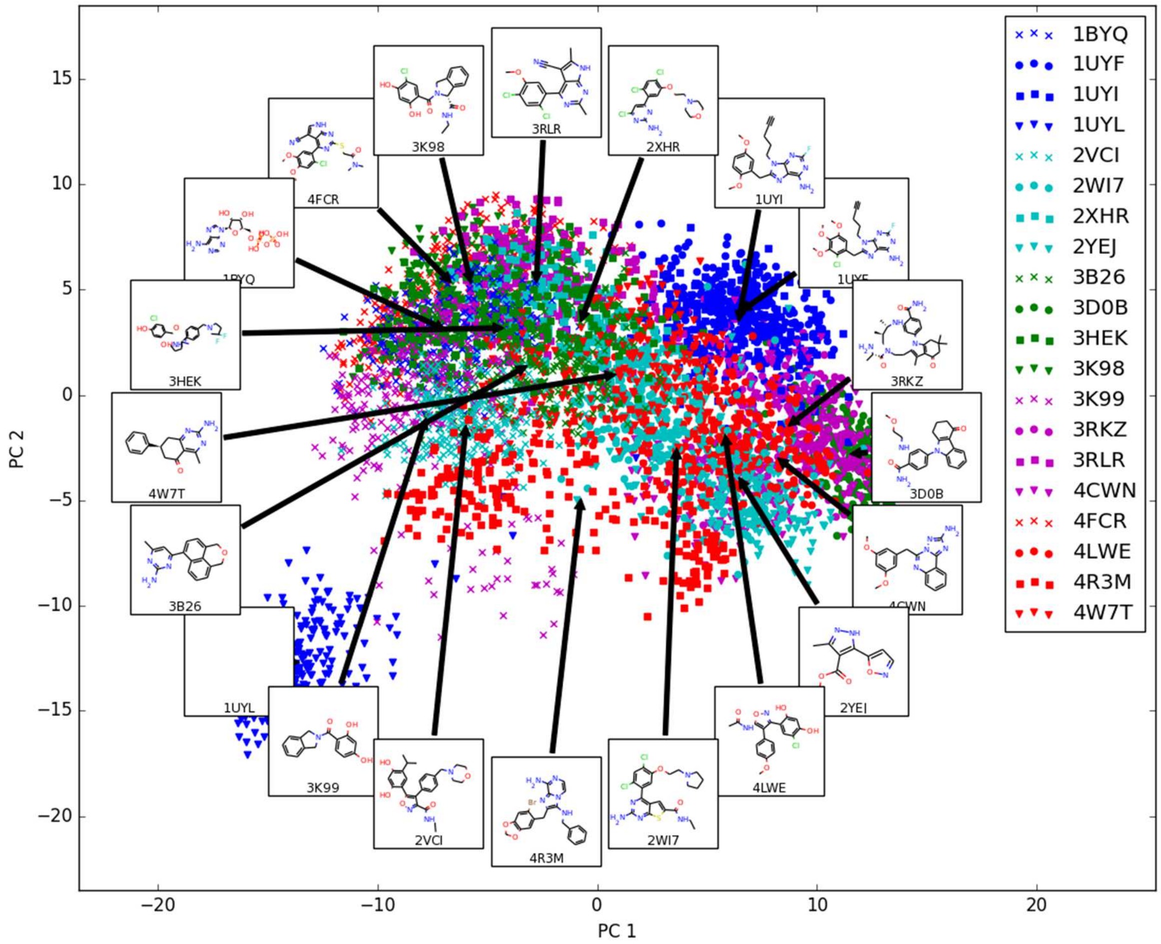


Figure 2.9 : Evenly-sampled HSP90 frames are overlaid on Principal Components 1 and 2.

Each simulation is indicated by a figure of its bound ligand with an arrow pointing to the centroid of its frames in PC space.

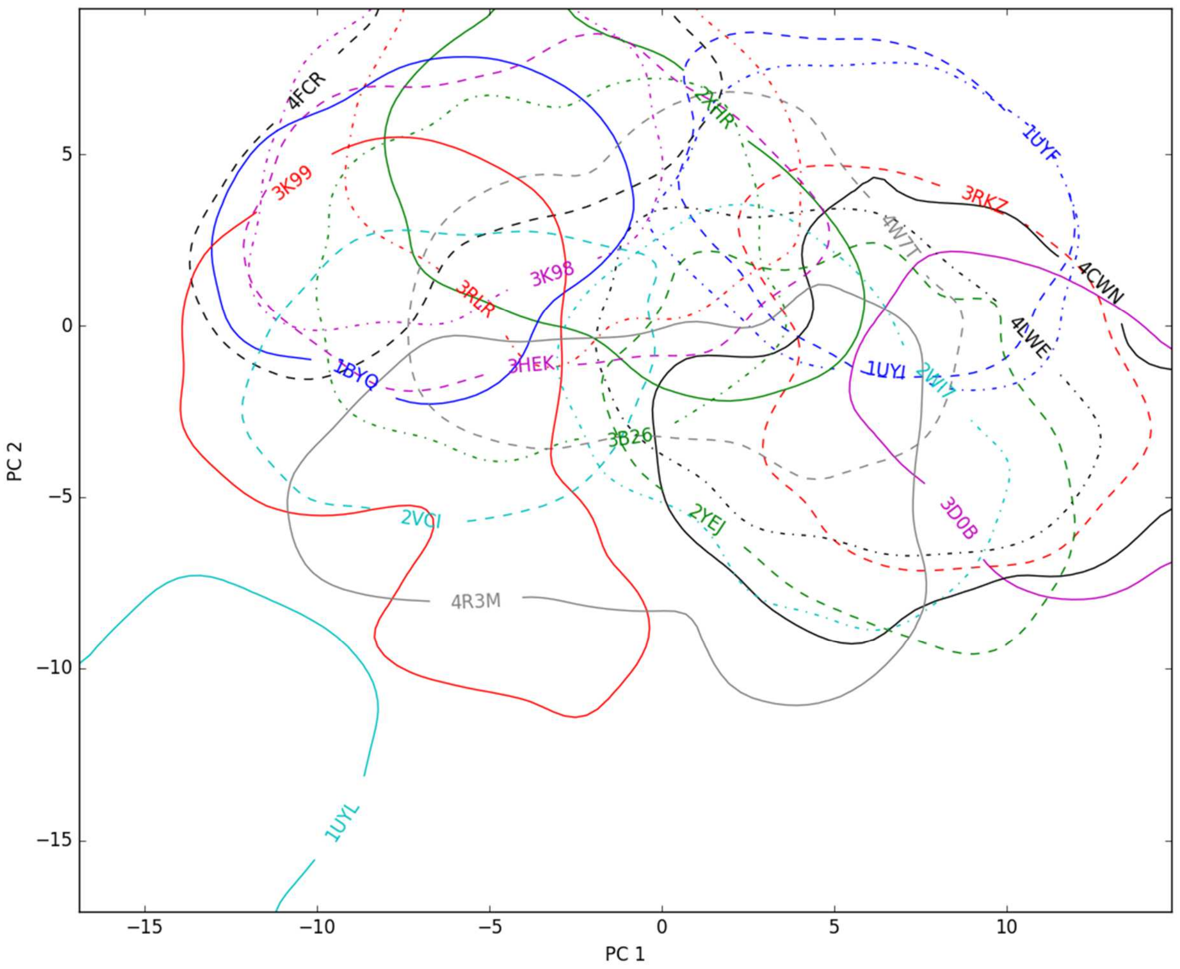


Figure 2.10 : Evenly-sampled HSP90 frames are overlaid on Principal Components 1 and 2, shown as a contour plot.

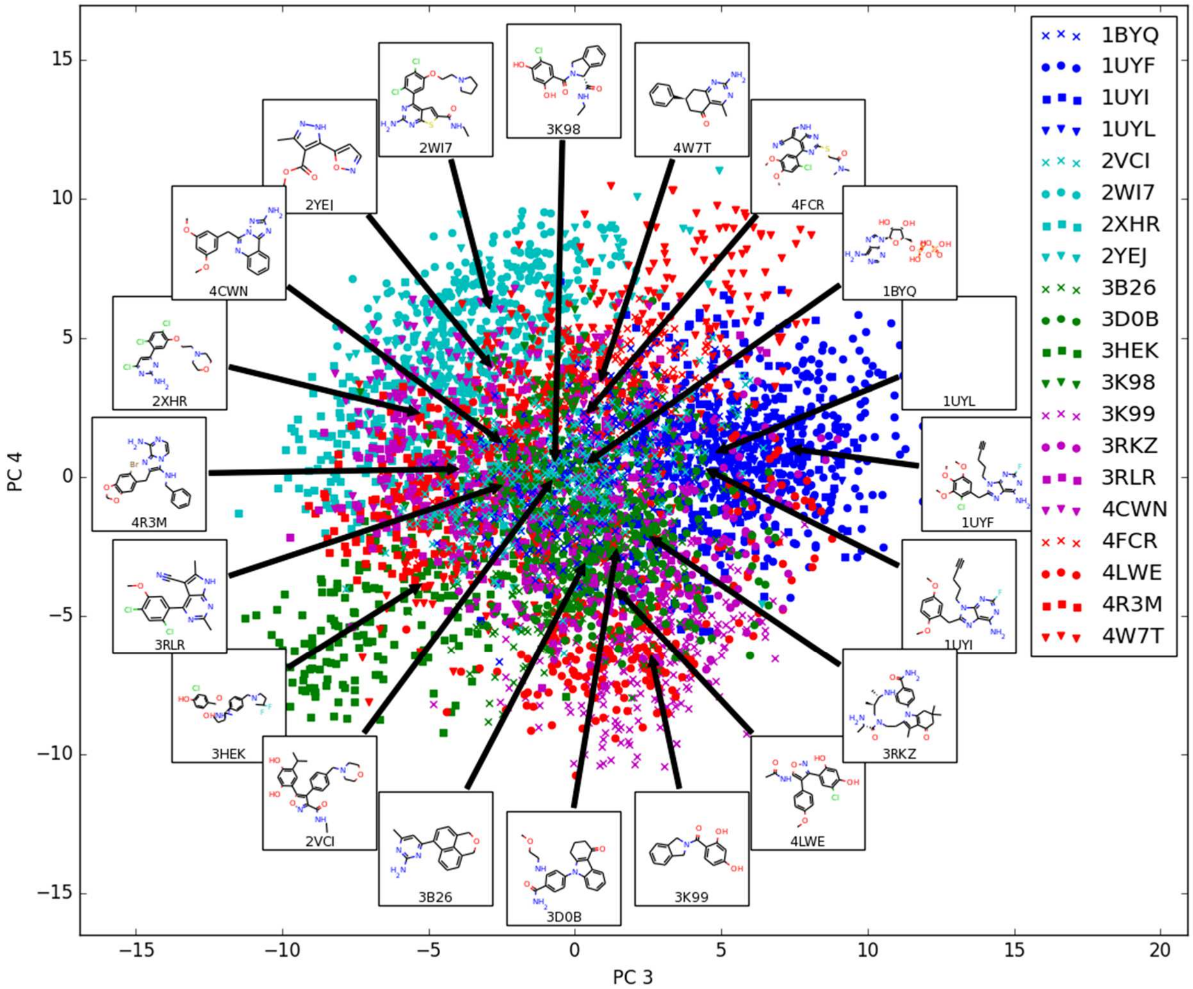


Figure 2.11 : Evenly-sampled HSP90 frames are overlaid on Principal Components 3 and 4.

Each simulation is indicated by a figure of its bound ligand with an arrow pointing to the centroid of its frames in PC space.

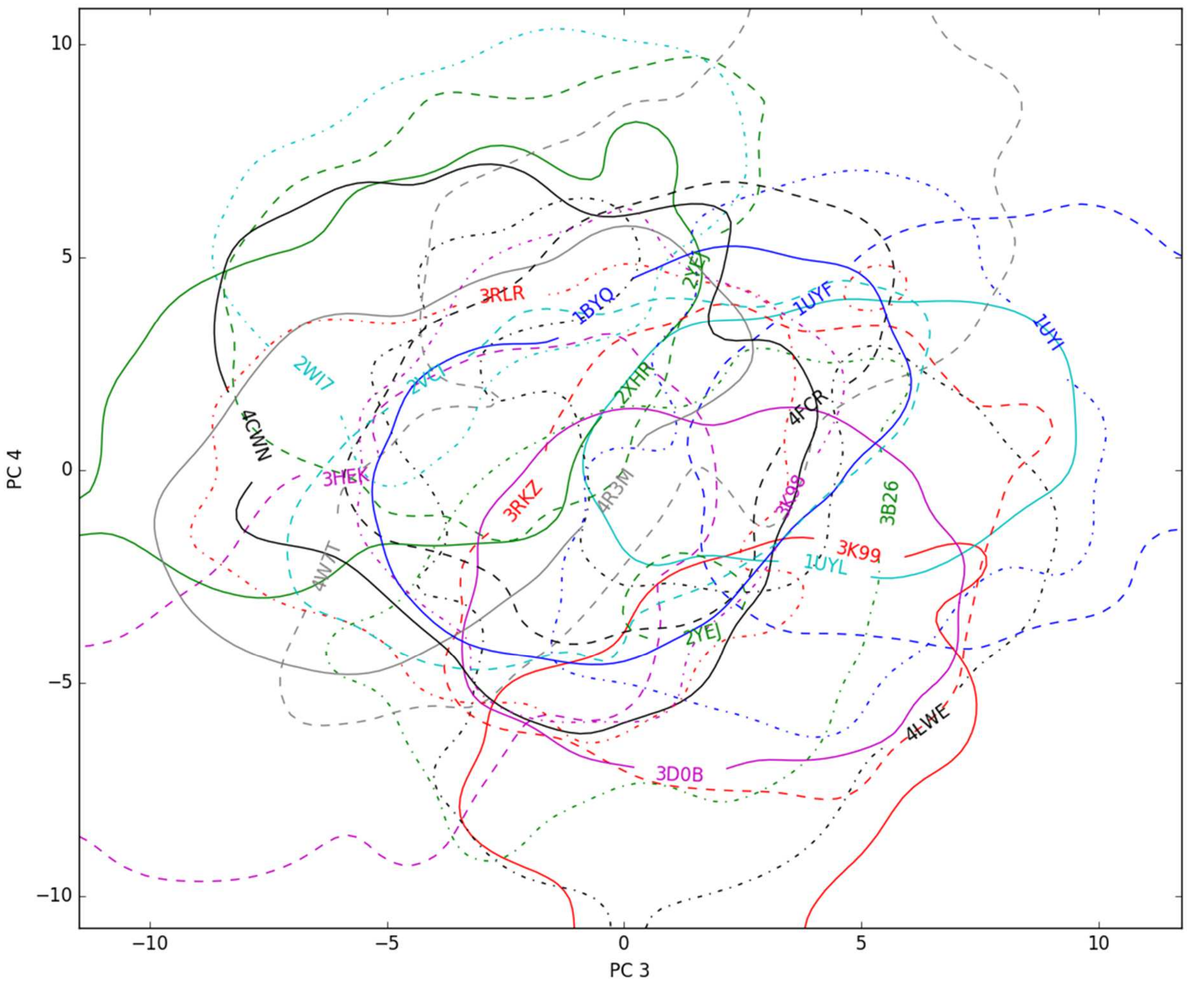


Figure 2.12 : Evenly-sampled HSP90 frames are overlaid on Principal Components 3 and 4, shown as a contour plot.

```
$SCHRODINGER/utilities/prepwizard -keepfarwat -disulfides -fillsidechains -fillloops -mse -
metal_binding -samplewater -propka_pH 7 -label_pkas $PDBID $PDBID_prepped.pdb -
reference_pdbid 1BYQ -LOCAL
```

An example prep command using Schrodinger Protein Prep Wizard

AMBER Input Scripts

S01-Min01-Proton.in

```
Minimization 01 - Proton
 &cntrl
   imin = 1,             ! Minimization (Yes)
   ntmin = 1,            ! Minimization Method (Steepest descent/Conjugate gradient)
   maxcyc = 2000,        ! Maximum number of minimization cycles (2000 cycles)
   ncyc = 1000,         ! Cycle of switch from steepest descent to conjugate gradient (at 
cycle 1000)
   cut = 10,            ! Non-bonding Cut-off (10 A)
   ntb = 1,             ! Periodic Conditions (Yes)
   ntr = 1,             ! Harmonic constraints in Cartesian space (Yes)
   restraint_wt = 10.0  ! Positional restraints weight ( 10 kcal/mol-A^2)
   restraintmask = "!@H=",  ! Restrained atoms (Not protons)
 /
```

S02-Min02-Solvent.in

```
Minimization 02 - Solvent
 &cntrl
   imin = 1,                ! Minimization (Yes)
   ntmin = 1,               ! Minimization Method (Steepest descent/Conjugate gradient)
   maxcyc = 2000,           ! Maximum number of minimization cycles (2000 cycles)
   ncyc = 1000,             ! Cycle of switch from steepest descent to conjugate gradient (at 
cycle 1000)
   cut = 10,                ! Non-bonding Cut-off (10 A)
   ntb = 1,                 ! Periodic Conditions (Yes)
   ntr = 1,                 ! Harmonic constraints in Cartesian space (On)
   restraint_wt = 10.0,     ! Positional restraints weight ( 10 kcal/mol-A^2)
   restraintmask = ":1-213 & :adp",      ! Restrained atoms (protein and ligand)
 /
```

S03-Min03-Focused.in

```
Minimization 03 - Focused
 &cntrl
   imin = 1,                ! Minimization (Yes)
   ntmin = 1,               ! Minimization Method (Steepest descent/Conjugate gradient)
   maxcyc = 2000,           ! Maximum number of minimization cycles (2000 cycles)
   ncyc = 1000,             ! Cycle of switch from steepest descent to conjugate gradient (at 
cycle 1000)
   cut = 10,                ! Non-bonding Cut-off (10 A)
   ntb = 1,                 ! Periodic Conditions (Yes)
   ntr = 1,                 ! Harmonic constraints in Cartesian space (On)
   restraint_wt = 10.0,     ! Positional restraints weight ( 10 kcal/mol-A^2)
   restraintmask = ":1-213",       ! Restrained atoms (protein)
 /
```

S04-Min04-Sidechains.in

```
Minimization 04 - Sidechains and Solvent
 &cntrl
   imin = 1,                ! Minimization (Yes)
   ntmin = 1,               ! Minimization Method (Steepest descent/Conjugate gradient)
   maxcyc = 2000,           ! Maximum number of minimization cycles (2000 cycles)
   ncyc = 1000,             ! Cycle of switch from steepest descent to conjugate gradient (at 
cycle 1000)
   cut = 10,                ! Non-bonding Cut-off (10 A)
   ntb = 1,                 ! Periodic Conditions (Yes)
   ntr = 1,                 ! Harmonic constraints in Cartesian space (On)
   restraint_wt = 10.0,     ! Positional restraints weight ( 10 kcal/mol-A^2)
   restraintmask = ":1-213@CA,N,C,O",     ! Restrained atoms (protein backbone)
 /
```

S05-Min05-All.in

```
Minimization 05 - All Atoms
 &cntrl
   imin = 1,                ! Minimization (Yes)
   ntmin = 1,               ! Minimization Method (Steepest descent/Conjugate gradient)
   maxcyc = 5000,           ! Maximum number of minimization cycles (5000 cycles)
   ncyc = 1000,             ! Cycle of switch from steepest descent to conjugate gradient (at 
cycle 1000)
   cut = 10,                ! Non-bonding Cut-off (10 A)
   ntb = 1,                 ! Periodic Conditions (Yes)
 /
```

S06-Eql01-Heating-NTV.in

```
Restrained Heating 250 ps NVT MD
 &cntrl
   ig = -1,                 ! Pseudo-random number generator (random seed based on time)
   irest = 0,               ! Restart the Simulation? (No)
   ntx = 1,                 ! Read in only initial coordinates (ASCII)
   cut = 10,                ! Non-bonding Cut-off (10 A)
   ntc = 2,                 ! SHAKE bond length constraints (constrain bonds with H)
   ntf = 2,                 ! SHAKE force evaluation (omit bonds with H)
                            ! Note: SHAKE set for TIP-type waters (e.g. TIP3P)
   ntb = 1,                 ! PBC (Constant Volume)
   ntt = 3,                 ! Temperature scaling (Langevin dynamics)
   gamma_ln = 1.0,          ! Collision frequency (1 ps^-1)
   tempi = 0.0,             ! Initial temperature (O K, velocities assigned according to 
forces)
   temp0 = 100.0,           ! Reference temperature (100 K)
   ntr = 1,                 ! Harmonic constraints in Cartesian space (On)
   restraint_wt = 5.0,      ! Positional restraints weight ( 5 kcal/mol-A^2)
   restraintmask = ":1-213@CA,N,C,O",      ! Restrained atoms (protein backbone)
   dt = 0.002,              ! Simulation time-step (0.002 ps or 2 fs)
   nstlim = 25000,          ! Simulation length (25000 steps or 50 ps)
   ntpr = 1000,             ! Energy save interval (every 1000 steps or 2 ps)
   ntwx = 5000,             ! Coordinate/trajectory save interval (every 5000 steps or 10 ps)
   ntwr = 25000,            ! Restart file only at end of run.
   iwrap = 1,               ! Coordinates to be "wrapped" into primary box (on)
   ioutfm = 1,              ! Trajectory file format (Binary NetCDF)
   nmropt = 1,              ! Turn on NMR restraints - so we can control temp increase (see 
below).
 /
 &wt type = 'TEMP0',        ! Variable Conditions Type (Temp)
   istep1 = 0,              ! Start Change Step (0)
   istep2 = 25000,          ! Last Change Step (25000 steps or 50 ps)
   imult = 0                ! Interplation (Linear (Default))
   value1 = 0.0,            ! Start State (0 K)
   value2 = 100.0 /         ! End State (100 K)
 &wt type='END' /
```

S07-Eql02-Heating-NTP.in

```
Restrained Heating 250 ps NVT MD
 &cntrl
   ig = -1,                 ! Pseudo-random number generator (random seed based on time)
   irest = 1,               ! Restart the Simulation? (Yes)
   ntx = 5,                 ! Read coordinates, velocities, and box
   cut = 10,                ! Non-bonding Cut-off (10 A)
   ntc = 2,                 ! SHAKE bond length constraints (constrain bonds with H)
   ntf = 2,                 ! SHAKE force evaluation (omit bonds with H)
                            ! Note: SHAKE set for TIP-type waters (e.g. TIP3P)
   ntb = 2,                 ! PBC (Constant Pressure)
   ntp = 1,                 ! Constant Pressure MD (Isotropic position scaling)
   barostat = 1,            ! Berendsen Barostat used for equilibration
   pres0 = 1.0,             ! Reference Pressure (1 bar)
   taup = 5.0,              ! Pressure relaxation time (5 ps)
   ntt = 3,                 ! Temperature scaling (Langevin dynamics)
   gamma_ln = 1.0,          ! Collision frequency (1 ps^-1)
   tempi = 100.0,           ! Initial temperature
   temp0 = 300.0,           ! Reference temperature (300 K)
   ntr = 1,                 ! Harmonic constraints in Cartesian space (On)
   restraint_wt = 5.0,      ! Positional restraints weight ( 5 kcal/mol-A^2)
   restraintmask = ":1-213@CA,N,C,O",     ! Restrained atoms (protein backbone)
   dt = 0.002,              ! Simulation time-step (0.002 ps or 2 fs)
   nstlim = 100000,         ! Simulation length (100000 steps or 200 ps)
   ntpr = 1000,             ! Energy save interval (every 1000 steps or 2 ps)
   ntwx = 5000,             ! Coordinate/trajectory save interval (every 5000 steps or 10 ps)
   ntwr = 100000,           ! Restart file only at end of run.
   iwrap = 1,               ! Coordinates to be "wrapped" into primary box (on)
   ioutfm = 1,              ! Trajectory file format (Binary NetCDF)
   nmropt = 1,              ! Turn on NMR restraints - so we can control temp increase (see
below).
 /
 &wt type = 'TEMP0',        ! Variable Conditions Type (Temp)
   istep1 = 0,              ! Start Change Step (0)
   istep2 = 75000,          ! Last Change Step (75000 steps or 150 ps)
   imult = 0                ! Interplation (Linear (Default))
   value1 = 100.0,          ! Start State (100 K)
   value2 = 300.0 /         ! End State (300 K)
 &wt type='END' /
```

S08-Eql03-EqlOnlyStage01.in

```
Restrained Equilibration Stage 1 250 ps NPT MD
 &cntrl
   ig = -1,                 ! Pseudo-random number generator (random seed based on time)
   irest = 1,               ! Restart the Simulation? (Yes)
   ntx = 5,                 ! Read coordinates, velocities, and box
   cut = 10,                ! Non-bonding Cut-off (10 A)
   ntc = 2,                 ! SHAKE bond length constraints (constrain bonds with H)
   ntf = 2,                 ! SHAKE force evaluation (omit bonds with H)
                            ! Note: SHAKE set for TIP-type waters (e.g. TIP3P)
   ntb = 2,                 ! PBC (Constant Pressure)
   ntp = 1,                 ! Constant Pressure MD (Isotropic position scaling)
   ntp = 1,                 ! Constant Pressure MD (Isotropic position scaling)
   barostat = 1,            ! Berendsen Barostat used for equilibration
   pres0 = 1.0,             ! Reference Pressure (1 bar)
   taup = 5.0,              ! Pressure relaxation time (2 ps)
   ntt = 3,                 ! Temperature scaling (Langevin thermostat) - Gives real dynamics
   gamma_ln = 5.0,          ! Collision frequency (5 ps^-1)
   temp0 = 300.0,           ! Reference temperature (300 K)
   ntr = 1,                 ! Harmonic constraints in Cartesian space (On)
   restraint_wt = 5.0,      ! Positional restraints weight ( 5 kcal/mol-A^2)
   restraintmask = ":1-213@CA,N,C,O",     ! Restrained atoms (protein backbone)
   dt = 0.002,              ! Simulation time-step (0.002 ps or 2 fs)
   nstlim = 125000,         ! Simulation length (125000 steps or 250 ps)
   ntpr = 1000,             ! Energy save interval (every 1000 steps of 2 ps)
   ntwx = 5000,             ! Coordinate/trajectory save interval (every 5000 steps of 10 ps)
   ntwr = 125000,           ! Restart file only at end of run.
   iwrap = 1,               ! Coordinates to be "wrapped" into primary box (on)
   ioutfm = 1,              ! Trajectory file format (Binary NetCDF)
 /
```

S09-Eql04-EqlOnlyStage02.in

```
Unrestrained Equilibration Stage 2 500 ps NPT MD
 &cntrl
   ig = -1,                 ! Pseudo-random number generator (random seed based on time)
   irest = 1,               ! Restart the Simulation? (Yes)
   ntx = 5,                 ! Read coordinates, velocities, and box
   cut = 10,                ! Non-bonding Cut-off (10 A)
   ntc = 2,                 ! SHAKE bond length constraints (constrain bonds with H)
   ntf = 2,                 ! SHAKE force evaluation (omit bonds with H)
                            ! Note: SHAKE set for TIP-type waters (e.g. TIP3P)
   ntb = 2,                 ! PBC (Constant Pressure)
   ntp = 1,                 ! Constant Pressure MD (Isotropic position scaling)
   barostat = 2,            ! Monte Carlo Barostat - Optimal for GPU runs
   mcbarint = 1000,         ! Steps between volume changes for the barostat
   pres0 = 1.0,             ! Reference Pressure (1 bar)
   taup = 2.0,              ! Pressure relaxation time (2 ps)
   ntt = 3,                 ! Temperature scaling (Langevin thermostat) - Gives real dynamics
   gamma_ln = 5.0,          ! Collision frequency (5 ps^-1)
   temp0 = 300.0,           ! Reference temperature (300 K)
   dt = 0.002,              ! Simulation time-step (0.002 ps or 2 fs)
   nstlim = 250000,         ! Simulation length (250000 steps or 250 ps)
   ntpr = 5000,             ! Energy save interval (every 5000 steps of 10 ps)
   ntwx = 5000,             ! Coordinate/trajectory save interval (every 5000 steps of 10 ps)
   ntwr = 250000,           ! Restart file only at end of run.
   iwrap = 1,               ! Coordinates to be "wrapped" into primary box (on)
   ioutfm = 1,              ! Trajectory file format (Binary NetCDF)
 /
```

S10-Pro01-MD_10ns.in

```
10 ns NTP MD
 &cntrl
   ig = -1,                 ! Pseudo-random number generator (random seed based on time)
   irest = 1,               ! Restart the Simulation? (Yes)
   ntx = 5,                 ! Read coordinates, velocities (ASCII)
   cut = 10,                ! Non-bonding Cut-off (10 A)
   ntc = 2,                 ! SHAKE bond length constraints (constrain bonds with H)
   ntf = 2,                 ! SHAKE force evaluation (omit bonds with H)
                            ! Note: SHAKE set for TIP-type waters (e.g. TIP3P)
   ntb=2,                   ! PBC (Constant Pressure)
   ntp = 1,                 ! Constant Pressure MD (Isotropic position scaling)
   barostat = 2,            ! Monte Carlo Barostat - Optimal for GPU runs
   mcbarint = 1000,         ! Steps between volume changes for the barostat
   pres0 = 1.0,             ! Reference Pressure (1 bar)
   taup = 2.0,              ! Pressure relaxation time (2 ps)
   ntt = 3,                 ! Temperature scaling (Langevin thermostat) - Gives real dynamics
   gamma_ln = 5.0,          ! Collision frequency (5 ps^-1)
   temp0 = 300.0,           ! Reference temperature (300 K)
   dt = 0.002,              ! Simulation time-step (0.002 ps or 2 fs)
   nstlim = 5000000,        ! Simulation length (5000000 steps or 10 ns)
   ntpr = 5000,             ! Energy save interval (every 5000 steps or 10 ps)
   ntwx = 5000,             ! Coordinate/trajectory save interval (every 5000 steps or 10 ps)
   ntwr = 5000,             ! Restart file save interval (every 5000 steps or 10 ps)
   iwrap = 1,               ! Coordinates to be "wrapped" into primary box (on)
   ioutfm = 1,            ! Trajectory file format (Binary NetCDF)
 /
```

S10-Pro01-MD_50ns.in

```
50 ns NTP MD
 &cntrl
   ig = -1,               ! Pseudo-random number generator (random seed based on time)
   irest = 1,             ! Restart the Simulation? (Yes)
   ntx = 5,               ! Read coordinates, velocities (ASCII)
   cut = 10,              ! Non-bonding Cut-off (10 A)
   ntc = 2,               ! SHAKE bond length constraints (constrain bonds with H)
   ntf = 2,               ! SHAKE force evaluation (omit bonds with H)
                          ! Note: SHAKE set for TIP-type waters (e.g. TIP3P)
   ntb=2,                 ! PBC (Constant Pressure)
   ntp = 1,               ! Constant Pressure MD (Isotropic position scaling)
   barostat = 2,          ! Monte Carlo Barostat - Optimal for GPU runs
   mcbarint = 1000,       ! Steps between volume changes for the barostat
   pres0 = 1.0,           ! Reference Pressure (1 bar)
   taup = 2.0,            ! Pressure relaxation time (2 ps)
   ntt = 3,               ! Temperature scaling (Langevin thermostat) - Gives real dynamics
   gamma_ln = 5.0,        ! Collision frequency (5 ps^-1)
   temp0 = 300.0,         ! Reference temperature (300 K)
   dt = 0.002,            ! Simulation time-step (0.002 ps or 2 fs)
   nstlim = 25000000,     ! Simulation length (25000000 steps or 50 ns)
   ntpr = 5000,           ! Energy save interval (every 5000 steps or 10 ps)
   ntwx = 5000,           ! Coordinate/trajectory save interval (every 5000 steps or 10 ps)
   ntwr = 5000,           ! Restart file save interval (every 5000 steps or 10 ps)
   iwrap = 1,             ! Coordinates to be "wrapped" into primary box (on)
   ioutfm = 1,            ! Trajectory file format (Binary NetCDF)
 /
```

Scripts used for MD simulation

Acknowledgements

Chapter 2 is a modified reprint of the material as it appears in "Wagner, J., Sørensen, J., Hensley, N., Wong, C., Zhu, C., Perison, T., Amaro, R.E. (2017) POVME3.0: Software for Mapping Binding Pocket Flexibility, *Journal of Chemical Theory and Computation*. doi: 10.1021/acs.jctc.7b00500". The dissertation author was the primary investigator and author of this paper.

This work was funded in part by the Director's New Innovator Award Program NIH DP2 OD007237 to REA. Funding and support from the National Biomedical Computation Resource

(NBCR) is provided through NIH P41 GM103426. JRW was supported by the NIH Molecular Biophysics Training Grant T32 GM008326. JS was supported by the Alfred Benzon Foundation. We thank Prof. Jacob Durrant, Dr. Lane Votapka, and Christopher Condon for helpful discussions.

Chapter 3 : Architecture and Dynamics of the ssDNA complex of wild-type APOBEC3B C-terminal domain

Jeffrey Wagner, Ozlem Demir, Rommie Amaro

Introduction

The APOBEC3 (A3) family of cytidine deaminases is a recently-discovered endogenous source of mutation in cancer[81]. Recent studies have linked cancer progression and recurrence to A3 expression levels. Many A3 protein have been found to show substrate sequence preferences, and analysis of some cancer genomes has shown an enrichment of A3 mutation signatures. Recently, efforts have begun to discover therapies for DNA damage caused by the A3 proteins that are responsible for the family's role in cancer. Evidence suggests that APOBEC3B (A3B) is the most important A3 family member in driving tumor progression.

Each A3 protein consists of either one or two deaminase domains. These domains share a common fold and a minimum sequence identity of 30%. Despite this high homology, in A3 proteins with two deaminase domains previous work has found that only the C-terminal domain (ctd) is catalytically active. Only single A3 domains have been solved through X-ray crystallography. Therefore, it is not known how the dual-domain A3 domains interface, or how the catalytically inactive N-terminal domain (ntd) affects protein function.

Interestingly, the substrate preferences of A3 proteins can be exchanged through the transfer of certain loops[82]. This discovery established the role of loops 1, 3, and 7 in the process of DNA substrate recognition. Initial crystal structures of A3 deaminase domains show these loops being adjacent to the binding site. Subsequent crystallization of A3 proteins in complex with optimized substrate sequences showed residues on loops 1 and 7 binding to oligonucleotide substrates. Previous experimental studies have explored the binding of cytidine deaminases to chemically modified oligonucleotides, such as those with a ribose-cytidine (rC) base at the target site, as well as other modifications to the oligonucleotide backbone and base[83]. These studies have concluded that A3B prefers DNA substrates, but can bind and catalyze other substrates with significantly lower activity. DNA and RNA differ in structure by one oxygen (should be hydroxyl group), but this difference has the effect of changing the preference for the backbone sugar ring pucker. It is suspected that this is a contributing factor for A3 proteins' preference for DNA over RNA, but the exact mechanism is not known.

One powerful technique to understand the biophysics of proteins and biological interactions is molecular dynamics (MD) computer simulations. These simulations model proteins starting in an initial configuration, and undergoing motion according to the laws of physics at physiologic temperature. With increases in computing power and the scope of computable questions, MD simulations have begun to find valuable synergies with traditional biochemistry, and can explain mechanisms underlying observations or propose new routes of experimentation. With recent developments, MD simulations have become capable of simulating not just proteins, but also numerous solvents, ions, nucleotides, and have general rules to parameterize small molecules.

In this work, we use molecular dynamics simulations to explore A3Bctd conformational dynamics and oligonucleotide recognition. We find that an analysis of these simulations reveals the importance of base-specific hydrogen bonds, pocket shape, and backbone sugar conformational preference in A3B substrate recognition. This correlation suggests that further observations from these models can be used to accelerate study of A3 proteins, and may generalize to other protein-oligonucleotide. Our findings regarding the biophysics of A3B advances our understanding of a major driver of mutation in cancer, and does so in a way that is directly applicable to drug design.

Methods

Simulations of A3B were parameterized using the AMBER FF14SB forcefield for protein atoms, and FF99BSC0 and FF99BSC0_chi0L3 force field for DNA and RNA atom, respectively. The starting coordinates for oligonucleotide-bound simulations were based on PDB entry 5TD5, and the apo simulation was started with coordinates taken from 5CQI and a separate one from 5TD5 by deleting the nucleotides. In each system, mutations were reverted to wild-type and missing residues were modeled using the Schrodinger PRIME software suite. Simulations were embedded in a TIP3P water box generated by LEaP from the AmberTools suite with a buffer distance of 10 A, with Na and Cl ions added to neutralize charge and attain a solvent concentration of 0.2M. Solvent in the crystal structure other than waters were removed. Crystal waters were left in place, and protonation states and hydrogen coordinates were assigned by VMD PropKa. The catalytic zinc ion and the zinc-chelating residues in the active site were modeled according to the Cationic Dummy Atom Model [84]. The catalytic Zinc was also modeled bound to a OH- ion, in order to model the pre-catalysis substrate recognition dynamics of A3B.

Four A3Bctd systems were simulated: A DNA-bound system based on coordinates from 5TD5 with nucleotide sequence TTCATG, a hybrid oligonucleotide-bound system based on coordinates from 5TD5 with nucleotide sequence TTrCATG (where rC indicates a ribonucleotide cytidine), an apo simulation based on coordinates from 5TD5, and an apo simulation based on coordinates from 5CQI. Each system underwent minimization in its forcefield, followed by gradual heating and equilibration with decreasing restraints. AMBER input scripts for each step are provided in the Supplemental Materials. Each apo system was simulated in triplicate and each oligonucleotide-bound system was simulated quintuplicate, differing in temperature initialization seed, and each replicate underwent 1 μs of unrestrained MD simulation in an NPT ensemble at 310 K.

Analysis of hydrogen bonds was performed using the MDTraj Python package[85], and visualized using Python's Matplotlib[86]. The existence of hydrogen bonds was defined by Baker-Hubbard criteria[87]. The hydrogen bond analysis was performed at increments of X ns in the trajectories. Only hydrogen bonds that appear in at least 15% of simulation snapshots are shown.

Pocket volumes were studied using POVME3.0[21-23], and visualized using Visual Molecular Dynamics[56]. The pocket region was defined by a set of inclusion spheres which cover the observed DNA-binding region. This region is defined as running between loops 1 and 7, down into the zinc-containing active site pocket, and out between loops 1 and 3. Because quantitative comparison of the pockets was performed, the POVME convex hull exclusion option was not used, per suggested POVME3.0 best practices[23]. All trajectories were aligned by their backbone atoms to the starting structure of the DNA-bound A3B MD simulation (after equilibration).

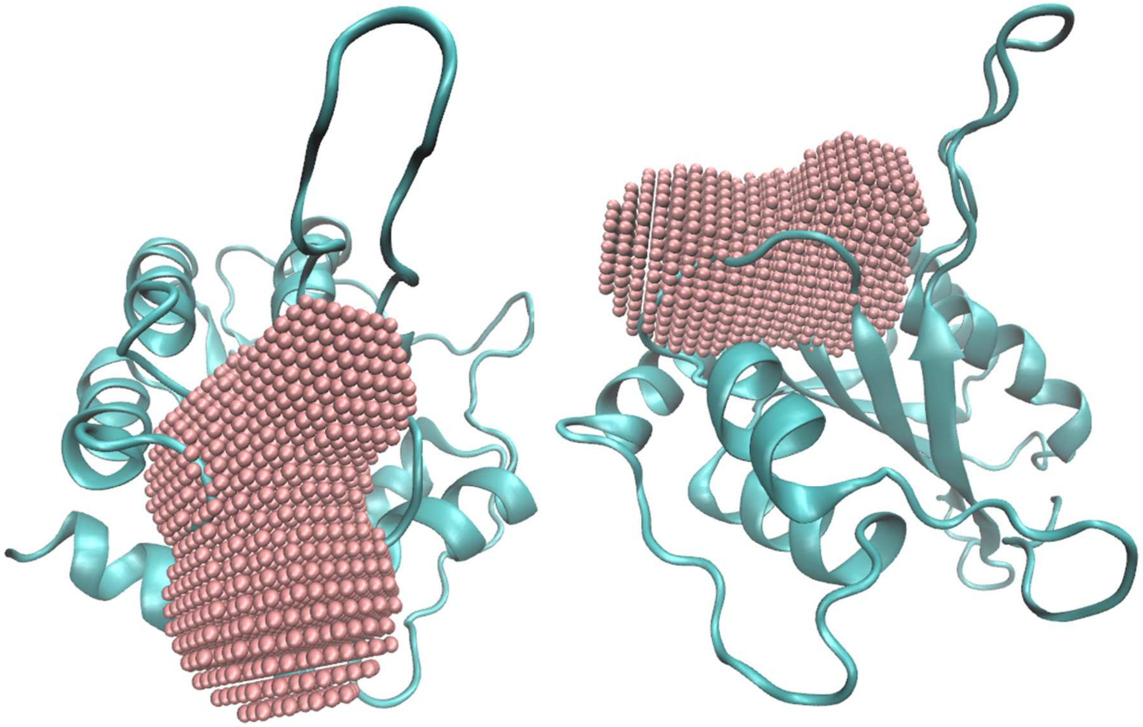


Figure 3.1: Inclusion Regions used for POVME analysis

Visualization of A3 dimer interfaces was performed using open-source PyMOL. The structures considered in this analysis were 5CQD[88], 5CQK[88], 5CQI[88], 5CQH[88], 4XXO[89], 2M65[90], 3VM8[91], 3VOW[91], 3WUS[92], 4J4J[93], 4IOU[94], 5HX4[95], 5HX5[95], 2MZZ[96], 5K81[97], 5K82[97], 5K83[97], 2JYW[98], 3E1U[99], 2KBO[100], 2KEM[101], 3IQS[99], 3IR2[102], 3V4K[103], 3V4J[103], 4ROV[104], and 4ROW[104]. Interface surface area was calculated using the EPPIC webserver[105]. Interfaces were filtered to only show those with more than 500 A^2 of surface area[106]. Structures which have a greater number of monomers in their asymmetric unit can show the same interface multiple times, and this is counteracted by only labeling each interface once.

Results and Discussion

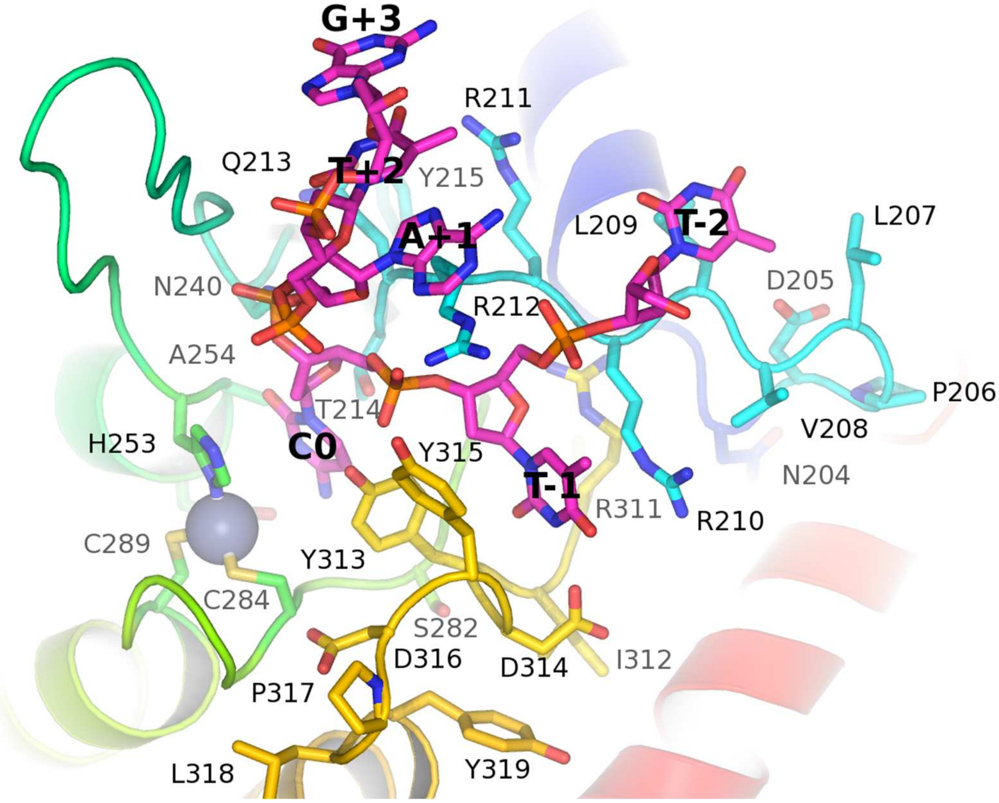


Figure 3.2 : Binding site-adjacent residues.

Differences between dC and rC target nucleotide simulations.

The simulations containing the ribose-C (rC) nucleotide at the target position displayed major differences from those containing the deoxyribose-C (dC). In the dC simulations, both the protein residues of the binding site and the -1, +1 and target cytidine nucleotides remained in the same position (RMSD < 2.2) throughout the simulations. However, in the rC simulations, the RMSDs of the same residues were much higher, as shown in Figure 3.1. This change indicates a shift in hybrid DNA binding pose, characterized by a binding site rearrangement. The simulations show that one of the the driving events in this shift was a change in the sugar pucker of the rC nucleotide (Figure 3.3).

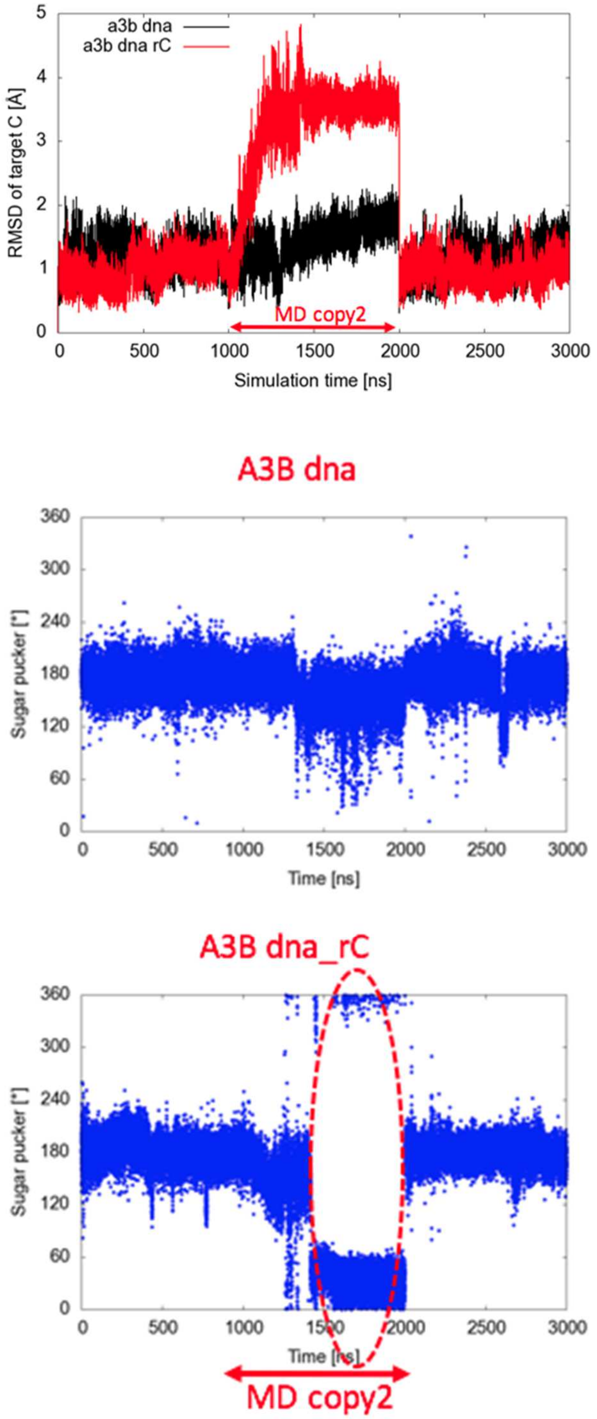


Figure 3.3 : RMSD of target C in A3B-DNA and A3B-hybrid DNA simulations(left). Sugar pucker of the target C, measured in both A3B-dC and A3B-rC simulations. In rC replicate 2, the RNA transitions from a C2’ endo to a C3’ endo conformation(right)

The rC shift also manifests as a change in binding site shape. Notably, the change in sugar pucker shifts the hybrid DNA away from its starting position and toward loop 1 and away from loop 3. The regions of pocket shape which are enriched in the hybrid DNA simulation are shown in red in Figure 3.4.

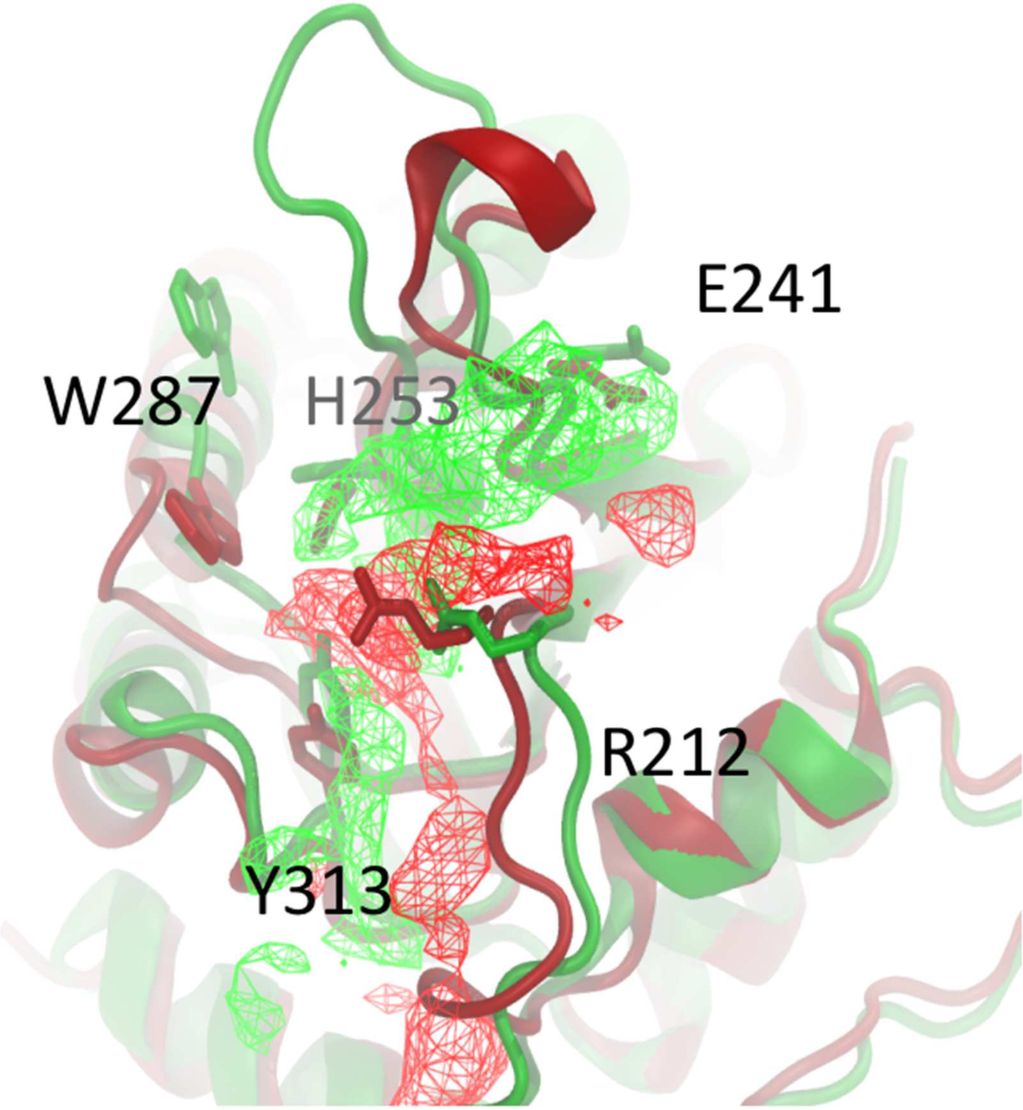


Figure 3.4 : Major regions of pocket shape difference observed between pure DNA (green) and hybrid DNA (red) simulations.

The shifted conformation can also be characterized by movements in the binding pocket residues, and changes in the network of hydrogen bonds made between the protein and target rC. Binding pocket residues which moved are shown in Figure 3.4.

After the shift, the 2’ O of the target rC forms a hydrogen bond with Thr214’s sidechain. In non-shifted simulations, the Thr214 sidechain maintains a hydrogen bond to the 4’ O of the target rC. This shift correlates with a breaking of all major protein-base hydrogen bonds in the other nucleotides in the chain, except that between Ser282 and the hydrolyzed amine on the rC. In all rC simulations, the initial hydrogen bond between the target C’s H41 and a Zinc-chelating residue, Glu255, is broken partway through each simulation. A table of all common hydrogen bonds between base and protein atoms is shown in Figure 3.5, and a complete bitmap is provided in the Supplemental Information.

Experiments on A3A have shown that Asp131, homologous to A3B’s Asp314, confers the preference for T at the -1 position of the oligonucleotide[107]. Both the crystal structures of A3B and our simulations show hydrogen bonds consistently formed between the H3 atom of the T -1 base. Interestingly, the hybrid simulation post-ring flip has broken this hydrogen bond, and replaced it with Asp316. In this portion of the simulation, the -1 T of the oligonucleotide makes a base hydrogen bond to the sidechain oxygen of Asp316. Asp316 has been shown to be essential for A3B antiviral function, and is therefore likely involved in DNA binding[108]. Given that this shift only appears in the hybrid DNA simulation, it is possible that the perturbation caused by the target C’s DNA-to-RNA mutation aided the simulation in leaving its initial energy well and exploring intermediate binding poses. It is also possible that Asp316 contributes by an indirect electrostatic mechanism when it is not directly involved in forming hydrogen bonds for substrate recognition.

While hydrogen bonding can therefore offer an explanation for the target C and T -1 base specificity, the simulations do not reveal specific sidechain-base hydrogen bonding for other nucleotides. The data implies, however, that shape-based recognition may take place. In our simulations, the positively charged sidechains of loop 1 residues contact the negatively charged phosphate backbone of the oligonucleotide. These positively charged loop 1 residues are known to be key for activity in A3A and A3Gctd, as A3A H29 and A3G H216(homologous to A3B R212) could be mutated to Arg while maintaining residual activity[109, 110]. However, when A3G H216 is mutated to Ala, it loses activity[98, 110]. While these backbone contacts appear to be charge-driven and are not specific to one nucleotide sequence, the base atoms of the nucleotide make consistent hydrogen bonds with the loop 1 backbone, in what may be a shape-driven recognition process.

Hydrogen Bond	TTCATG sim	TTrCATG pre ring flip	TTrCATG post ring flip
A +1 N3 - Arg212 sidechain guanidinium	21%	17%	42%
Target C O2' - Thr214 OH	N/A	3%	94%
Target C O2'H - His253 NE2	N/A	18%	0%
Target C O3' - Asn240 HD21	22%	42%	0%
Target C O4' - Thr214 OH	89%	74%	0%
Target C O2 - Ala254 backbone NH	95%	76%	0%
Target C H42 - Ser282 backbone C=O	96%	85%	93%
Target C H41 - Glu255 sidechain O	85%	35%	0%
T -1 H3 - Asp314 sidechain O	94%	86%	0%
T -1 H3 - Asp316 sidechain O	0%	0%	96%
T -1 O2 - Tyr 315 backbone NH	51%	59%	0%

Figure 3.5 : Protein-nucleobase hydrogen bonds which differ significantly between DNA and DNA-rC simulations.

After the shift of the hybrid DNA, the target cytosine is too far from the catalytic Zinc to perform deamination, and the catalytic residues are distorted from their crystal geometry. This new binding mode may be an intermediate conformation in normal DNA binding, however it was only observed in the rC simulation. Further simulation of all systems might eventually show the same shift, or complete oligonucleotide dissociation.

The oligonucleotide interaction surface for pure DNA and DNA-rC simulations are shown in Figure 3.6. Notably, the hybrid oligonucleotide explores more of the protein surface than the pure DNA oligonucleotide. While the target rC in the hybrid simulation remains flipped toward the protein, the surrounding nucleotides deviate from their positions in the pure DNA simulation. In the DNA-rC simulation, the flexible loop 3 maintains contact with the +2 and +3 residues of the oligonucleotide, whereas the -1 and 2 residues are no longer tightly held in the binding cleft.

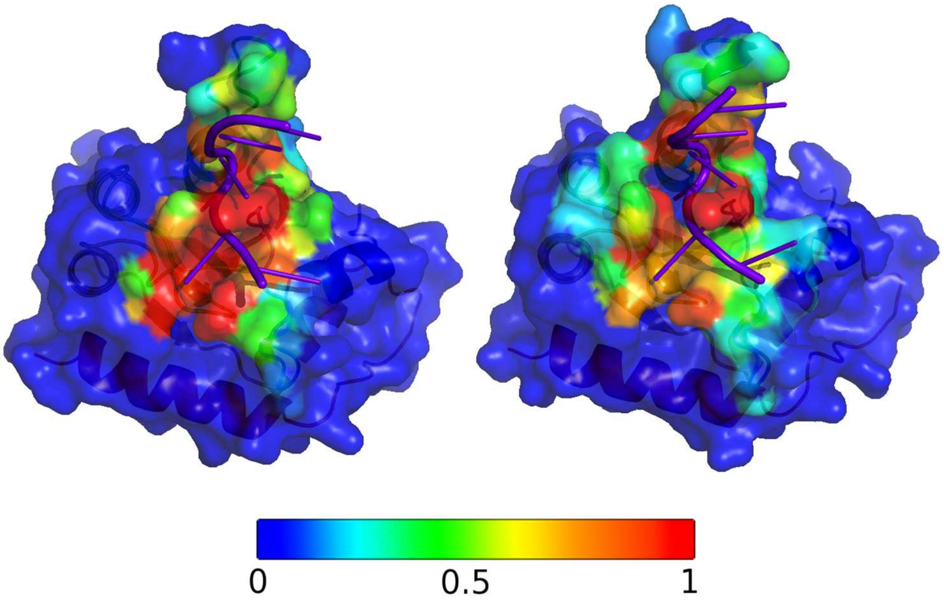


Figure 3.6 : Oligonucleotide interaction surface differences between pure DNA (left) and hybrid DNA/RNA (right) simulations.

The frequency of contact with the oligonucleotide atoms is shown for each protein atom on a color scale from blue (no contact) to red (frequent contact). The hybrid DNA/RNA simulation shows a less stable binding pose, in which a larger binding surface is explored, but some core binding residues are not consistently contacted.

Shifting patterns of loop-loop contacts

Figure 3.6 shows a summary of loop-loop contacts between the regions of A3B responsible for substrate recognition. The simulations show different patterns of loop-loop contacts, which may offer clues about substrate recognition mechanisms. Loops 1, 3, and 7 have been identified as being primarily responsible for substrate recognition, and our data indicate that their interaction patterns are heavily affected by the presence of nucleotides.

The apo simulations show extensive loop 1 - loop 3 interactions, specifically Arg212 and Gln213 to Asn240, Glu231, Ala242, and Lys243. These contacts are made less frequently in the

DNA-bound and hybrid post-ring flip simulations. This is to be expected, as the substrate passes directly between loops 1 and 3. The fact that the non-ring flipped rC simulations show an intermediate extent of loop1-3 contacts might be indicative of the poor fit of the modified substrate.

The apo simulations also show the most loop 1 - loop 7 contacts. This observation matches expectations, as the substrate oligonucleotide passes directly between these loops. Arg311 in loop 7 makes contact with most residues in the first half of loop 1, from Asn203 to Arg210. The apo simulation is the only in which Tyr313 contacts loop 1, primarily via Arg211, but also sometimes through the flanking Arg210 and 212. Both the apo and rC post-ring flip simulations show frequent contacts between Tyr315 on loop 7 and Pro206 to Arg210 on loop 1. Generally, the large number and frequency of contacts in the apo and hybrid post-ring flip simulations indicate a more closed binding site, again implying that the post-ring flip simulation may have captured an intermediate-bound state of the complex.

Both the DNA and hybrid DNA-bound simulations show significantly fewer loop contacts, commensurate with their high number of loop-substrate interactions.

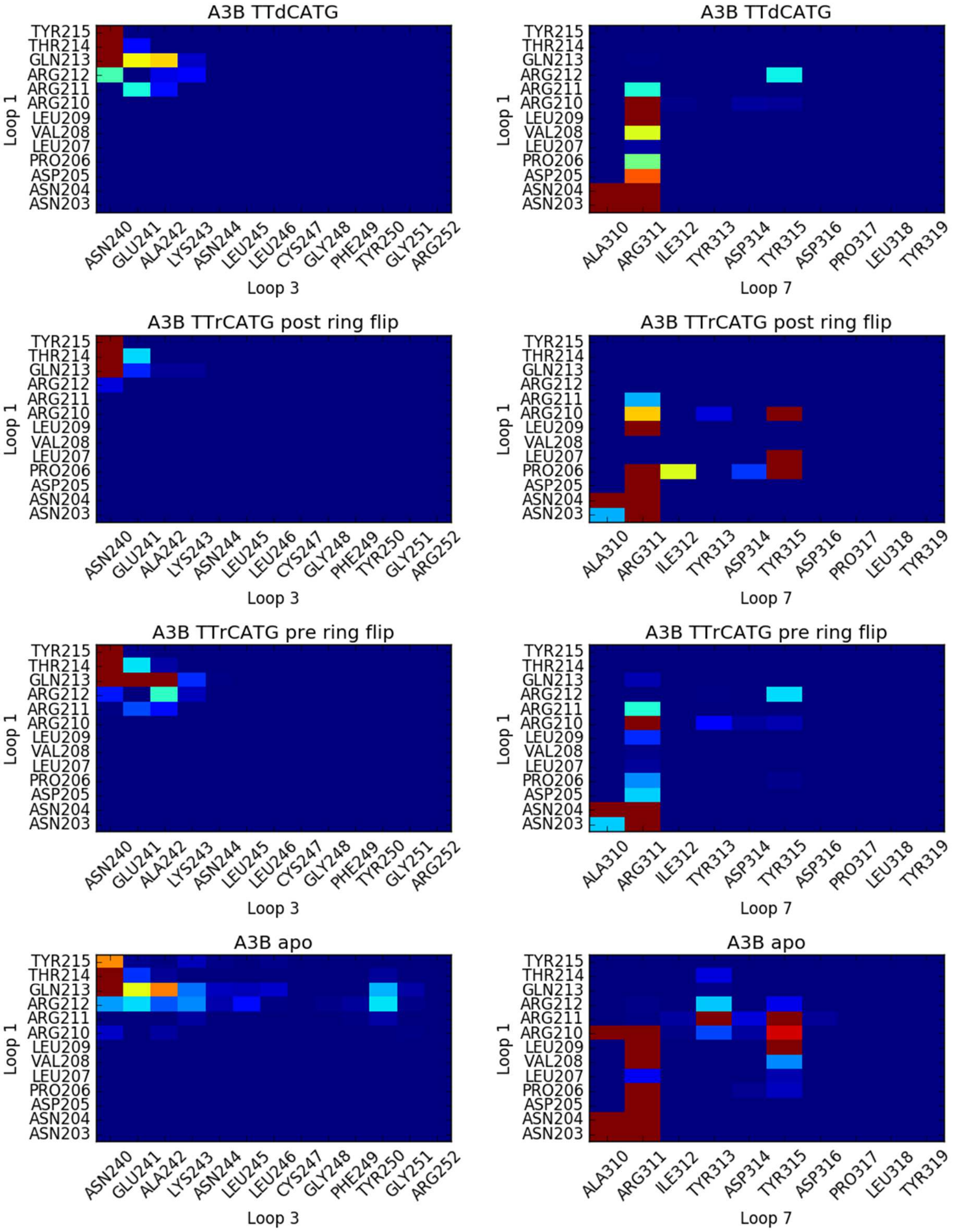


Figure 3.7 : Shifting patterns of loop-loop contacts in A3B simulations.

Blue indicates infrequent contacts, and dark red indicates contacts 25% of the time or more. Contacts are defined as a closest heavy atom distance of 4 Angstroms.

Analysis of crystal structures to discover oligomerization interfaces of A3 domains

Previous work has shown that interaction between A3 domains is an important phenomenon. Evidence for this interaction has been seen both in the activity differences of dual-domain APOBEC3s when expressed as full-length versus as the catalytic domain alone, and in the in-vitro oligomerization of wild type A3 domains which also frequently leads to activity differences. While no full-length A3 crystal structures have yet been solved, the packing of single-domain structures may offer hints to the basis of these observed oligomerizations. Figure 3.8 shows the frequently-observed crystal packing arrangements of A3 domains, which gives rise to two clusters of interfaces.

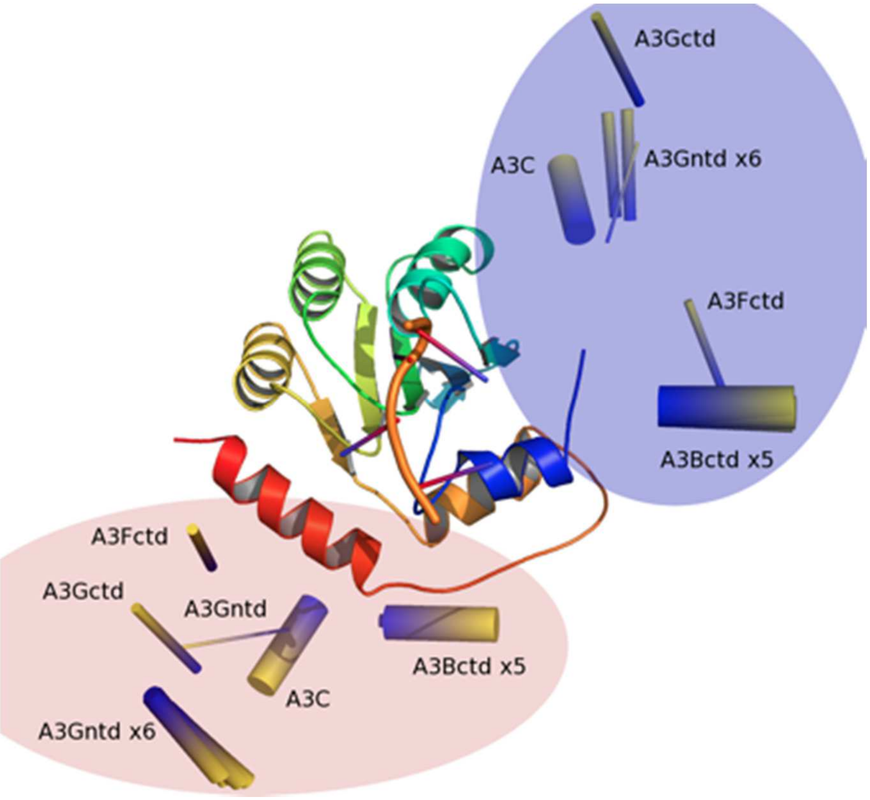


Figure 3.8 : Analysis of crystal structures to discover A3 domain oligomerization interface.

A3 crystal structures were analyzed to find crystal interfaces with at least 500 A^2 surface area. Each interface is shown according to the position of the other domain as a cylinder, with the yellow end at the center of mass of the domain and the blue end indicating the position of the catalytic Zn. The analysis shows clusters of interfaces at the N and C terminals of the reference domain (shown as blue and red cartoon/highlighting, respectively).

Comparative structural biology of APOBEC enzymes

Due to the high degree of homology between APOBEC3 enzymes, we expect that structural and functional insights from one family member may have implications for others. To this end, we have developed a Python package to explore structural and biochemical data within the APOBEC3 family. This package contains all APOBEC3 domains, pre-aligned to each other. It also enables annotation of structures, which can be visualized in Pymol as an overlay on crystal structures or homology models. This package not only shows annotations on the structure that they belong to, but can also identify the homologous residues that the annotations apply to on other APOBEC3 domains Figure 3.9 shows an example of two such aligned APOBEC3 domains and with an annotation of Asp314 in A3B and the homologous residue in A3G.

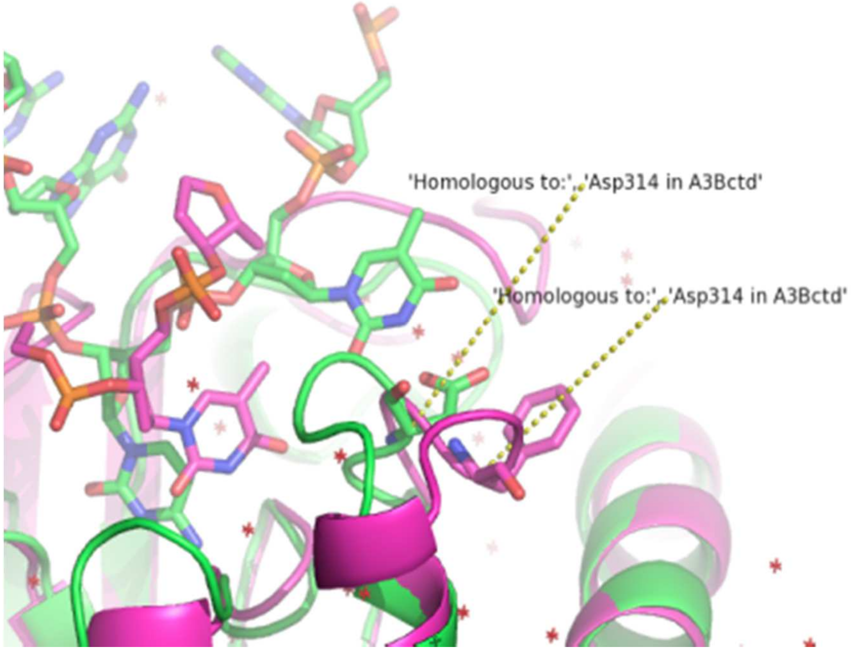

'Homologous to:', 'Asp314 in A3Bctd'
'Homologous to:', 'Asp314 in A3Bctd'

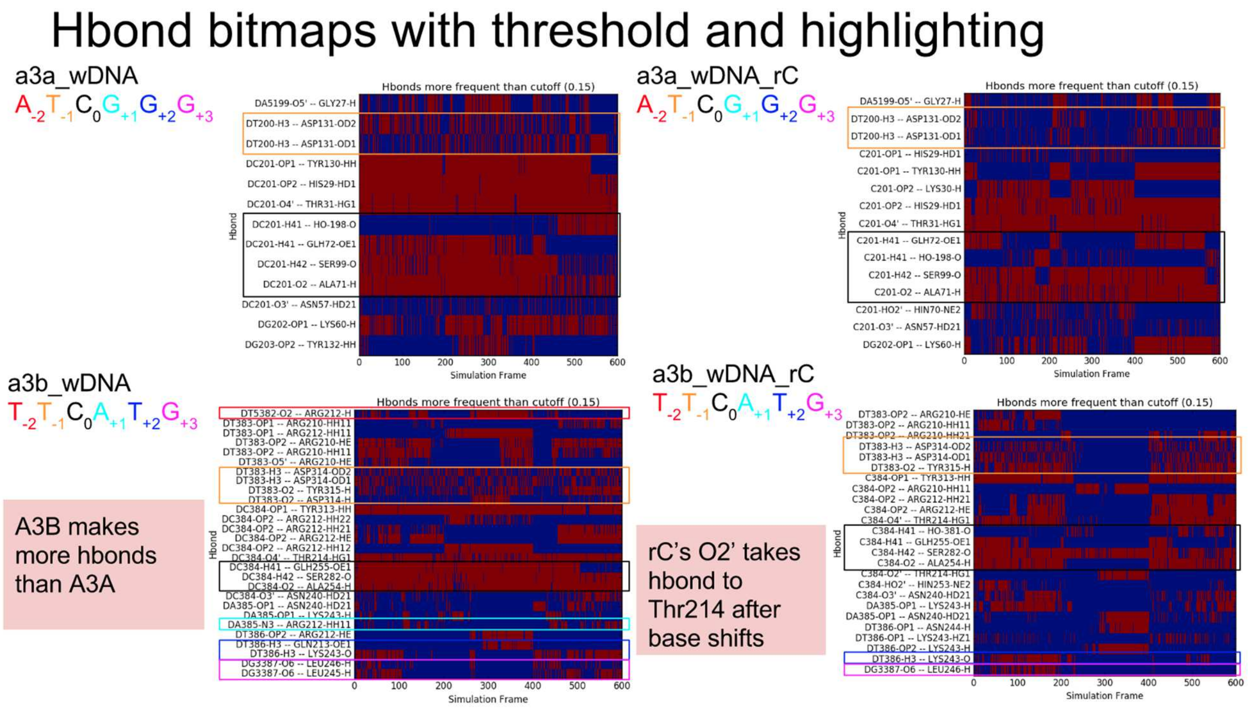
Figure 3.9 : Structural and functional annotation of APOBEC3B (green) and APOBEC3G (purple), with the important A3B substrate recognition residue Asp314 and its homologue in A3G annotated.

Conclusions

APOBEC3B is a recently-discovered source of mutation in cancer, and understanding the underlying biophysics of its activity has significant medical and scientific value. In this work, we revert crystal structures obtained using mutagenesis to the wild type sequence, and investigate the resulting protein dynamics using MD simulations. These simulations provide insight into potential mechanisms of substrate recognition and binding. Further, they show significant differences resulting from the presence of a substrate oligonucleotide.

Given the strong homology between A3 domains, it is likely that the understanding of one A3 protein can contribute knowledge to the study of others. To this end, we provide an analysis of crystal packing interfaces that might be useful toward to investigation of oligomerization and A3 domain interaction. We also provide a tool to easily compare A3 domain structures and the biochemistry of their sequence regions.

Additional Information



Hbond bitmaps with threshold and highlighting

A3B makes more hbonds than A3A

rC's O2' takes hbond to Thr214 after base shifts

Figure 3.10 : Full bitmaps of hydrogen bonds (threshold = 15% of frames) for A3A and A3B simulations

Acknowledgements

This work was funded in part by the Director's New Innovator Award Program NIH DP2 OD007237 to REA. Funding and support from the National Biomedical Computation Resource (NBCR) is provided through NIH P41 GM103426. JRW was supported by the NIH Molecular Biophysics Training Grant T32 GM008326.

Chapter 4 : Emerging Computational Methods for the Rational Discovery of Allosteric Drugs

Jeffrey R. Wagner, Christopher T. Lee (Contributed equally to this work), Jacob D. Durrant, Robert D. Malmstrom, Victoria A. Feher, Rommie E. Amaro

Abstract

Allosteric drug development holds promise for delivering medicines that are more selective and less toxic than those that target orthosteric sites. To date, the discovery of allosteric binding sites and lead compounds has been mostly serendipitous, through high-throughput screening. Over the last decade, structural data has become readily available for larger protein systems and more membrane protein classes (e.g., GPCRs and ion channels), which are common allosteric drug targets. In parallel, improved simulation methods now provide better atomistic understanding of the protein dynamics and cooperative motions that are critical to allosteric mechanisms. As a result of these advances, the field of predictive allosteric drug development is now on the cusp of a new era of rational structure-based computational methods. Here, we review algorithms that predict allosteric sites based on sequence data and molecular dynamics simulations, describe tools that assess the druggability of these pockets, and discuss how Markov state models and topology analyses provide insight into the relationship between protein dynamics and allosteric drug binding. In each section, we first provide an overview of the various method classes before describing relevant algorithms and software packages.

Review Motivation and Organization

To date, most allosteric drugs have been discovered through high-throughput screening. But growing databases of biomolecular structure and sequence data, in conjunction with increases in computing power and improvements in predictive algorithms, are enabling the rational de novo design of allosteric drugs. Given the large number of published algorithms for predicting allosteric mechanisms, it can be difficult to select the most appropriate method for a given target. This review serves as an introduction for those who wish to use computational techniques to develop allosteric drugs.

After a broad overview of allosteric drug discovery, this review is divided into three sections. First, we discuss bioinformatics and molecular-dynamics methods to identify allosterically important sequence positions. Second, we summarize the computational methods to predict druggable pockets at these functionally relevant sites. Finally, we describe how Markov state models and topological analyses can tie these single sequence sites to global protein function and dynamics.

Introduction

Allosteric drugs offer a number of advantages that make them desirable as drug candidates. Allosteric effectors, by definition, alter protein activity by binding to a site distinct from the orthosteric pocket. Because allosteric sites are typically less evolutionarily conserved, allosteric drugs can be highly selective, even among other members of the same protein family.[111-118] In some cases, allosteric sites are so unique among proteins that an effector is said to have “absolute subtype specificity."[112, 113, 119, 120]

Allosteric modulators may have spatiotemporal specificity. For example, they can be active only in the presence of the endogenous ligand, thus restricting their effect to certain tissues at certain times,[112-114, 119, 121] which may slow desensitization.[120, 122]

Allosteric effectors are generally saturable, meaning that they have a maximal effect that does not necessarily correspond to complete inhibition or activation.[112, 114-116, 118, 120, 121] This saturability enables safer dosing. For example, if the maximal effect is an 80% reduction in signaling, overdosing will not fully eliminate an essential signal.[111, 112, 114]

Other advantages can include noncompetitive inhibition (ie drug activity cannot be ”overwhelmed" by high concentrations of the endogenous ligand) and pathway- or substrate-specific modulation, which reduces unwanted activity by specifically targeting a single protein function.[112, 114, 120] For example, if a protein is involved in multiple pathways, an allosteric effector may impact the activity of each pathway differently depending on the systems-biology context. If a protein acts on multiple substrates, the impact on activity may depend on the biological context.

Despite many potential advantages of allosteric therapeutics, it has been challenging to identify predictive approaches to discovering allosteric drugs. In recent decades, the pharmaceutical industry has favored more traditional targets for three primary reasons: the relative ease of assay development around orthosteric sites; access to high-throughput, high-resolution X-ray crystallography; and advances in ligand- and receptor-based computational methods to optimize ligand-binding affinity at a substrate-competitive site. This structure-based approach is thought to significantly reduce the time and cost of hit-to-lead and lead-to-drug development by reducing the number of compounds that need be synthesized.[3, 123] Work by

Doman et al. comparing computer-aided drug discovery (CADD) and high-throughput screening (HTS) reported that the two methods had hit rates of 35% and 0.021%, respectively.[124]

In contrast, allosteric drugs are uniquely challenging from a rational drug-design perspective. Because experimental assays typically measure orthosteric function rather than ligand binding at the allosteric site, efficient development of allosteric drugs requires that the complex structure-activity relationships (SARs) governing both binding affinity and allosteric activity be considered simultaneously.[115, 120, 125] Further, allosteric sites are less likely to be evolutionarily conserved. While this enables increased subtype specificity, it also increases the chances of evolved resistance[115, 119, 120] and can complicate testing in evolutionarily distant animal models.[120]

Additionally, allosteric effectors are particularly susceptible to "mode switching," where relatively minor chemical changes can drastically affect ligand efficacy.**[120, 125]** Structurally similar drug metabolites, therefore, may have varying and unpredictable distributions and allosteric effects.**[120, 125]** Optimizing allosteric modes of action requires methods that are very different than those used in orthosteric drug discovery.**[115]**

Multifunctional allosteric proteins are particularly challenging. While drug designers may desire to target a single protein function, an allosteric effector may also alter other functions, hindering a full mechanistic understanding of the pharmacology.**[120]** Also, the benefits of spatiotemporal specificity are lost if the distribution of the endogenous ligand changes with progression of the disease state.**[120]** Finally, assessment of the limited number of known allosteric pockets indicates that they are generally shallow**[115]** and present flat SAR.**[120]** These structural features similarly challenge existing rational drug-discovery paradigms and the general practice of developing selective compounds by optimizing affinity.

Despite these challenges, allosteric drug discovery has gained momentum recently due to a number of developments.[120] First, several allosteric drugs across a broad range of pharmacological target classes have been rationally designed,[126-132] encouraging pursuit of others, as evidenced by the number of allosteric drugs currently in clinical trials.[133] The recent elucidation of new membrane protein crystal structures for GPCRs[134, 135] and ion channels[136, 137] have assisted in the structure-based design approach to these successes. Finally, advances in our understanding of allosteric mechanisms have supported development of additional rational design strategies (see below).

Our understanding of allosteric mechanisms has advanced considerably since the initial conception of Monod, Wyman, and Changeux.[138] Modern models of allostery consider conformational ensembles.[115, 118, 119, 139-152] This revised view supports newly established and emerging computational advances that comprehensively map conformational landscapes and predict communication between allosteric and orthosteric sites. For example, the physical mechanisms of allostery generally alter the entropic and enthalpic factors that define the conformational landscape and, therefore, govern protein function.[115, 118, 119] The observed correlation between allosteric modulation and protein structural dynamics is varied: Major conformational rearrangements occur in some cases, as compared with subtle shifts in conformational populations in others.[114-116, 119, 120, 144, 146, 153] An excellent metaphor for these phenomena are Kornev and Taylor's classification of "domino" versus "violin" models of allosteric signal transduction.[154]. Further, allosteric signals are transmitted through a range of structural motifs, from rigid core regions to flexible linkers.[155] Allostery may occur through essential residues along a single allosteric path[156] or through many weak pathways connecting one site to another, acting in concert.[157] As such, it is not surprising that many of the allostery-prediction methods discussed in this review (some of which do not use structure/geometry information at all) in practice may identify non-contiguous groups of residues as being allosterically linked. Such predictions should not immediately be assumed to be wrong but rather may indicate a non-"domino" model of allostery.

Recent work has also revealed that protein allostery is not merely a transition between two discrete protein conformations, as initially thought, but rather a shift in the equilibrium populations of many conformations, induced by effector binding.[140, 144, 147, 150, 151, 158, 159] It is becoming increasingly clear that the kinetics of these transitions define the mechanisms of allostery.[147, 158] Empowered with these new understandings and advances in molecular simulation (in terms of speed of calculation and improved methodologies), the era of allosteric drug discovery is now on the cusp of radical advancement.

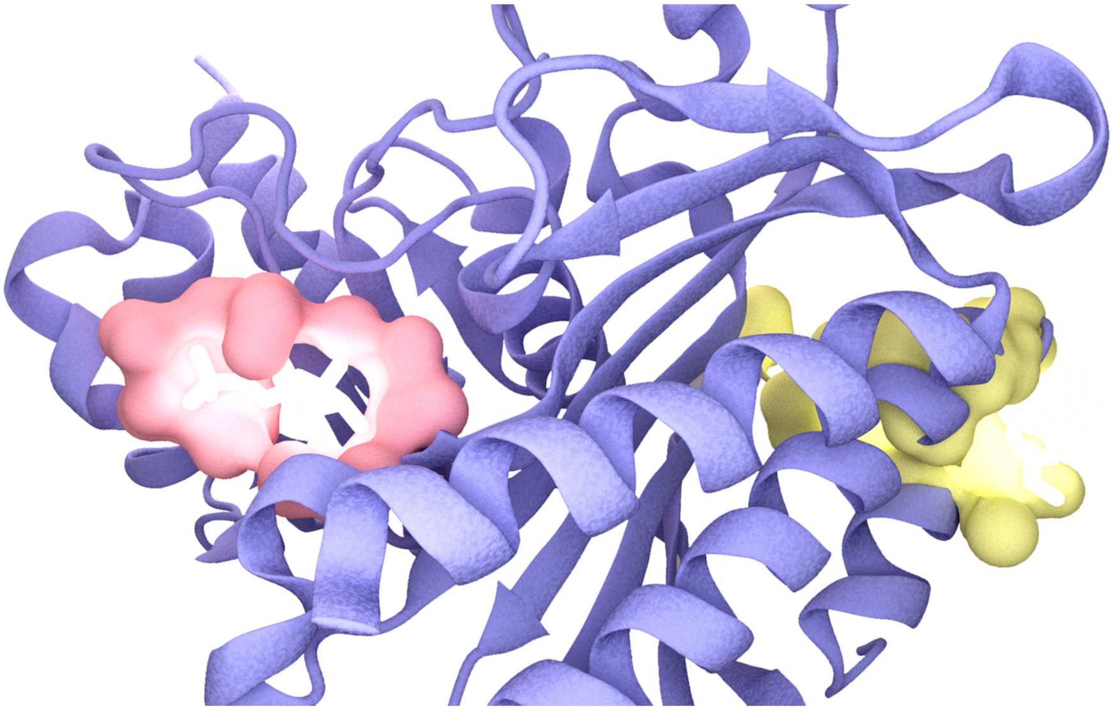


Figure 4.1 Orthosteric and Allosteric Sites.
The allosteric protein fructose 1,6-bisphosphatase, shown for illustration. Orthosteric and allosteric pockets (yellow and red, respectively) are bound to an endogenous ligand and an allosteric effector, respectively. Note that the allosteric site is distant from the orthosteric site such that there is no overlap between the bound poses of the allosteric and orthosteric ligands. Despite the distance between them, the allosteric effector measurably modifies the enzymatic activity at the orthosteric site. Illustration derived from PDB IDs 2Y5K[160] and 3IFC.[161]}

Emerging Rational Design Principles

So-called tried-and-true "design principles" are still being developed. However, a few general principles have begun to emerge. For example, many argue that it is insufficient to design a ligand that merely binds to an allosteric site; rather, the effector must make contact with certain key binding-pocket atoms to have the desired effect.[115, 120] These key atoms can often be identified through mutagenesis experiments and crystallographic studies of other allosteric ligands.

A promising set of design principles is encapsulated in the "allo-network" strategy, a rational approach that adopts two simultaneous but orthogonal approaches to ligand design.[116] On the protein-structure level, the primary focus is to target a single protein function or an interaction with a single partner. On the signaling-pathway level, the ”allo-network" strategy suggests targeting less-connected upstream proteins instead of the more direct, though potentially highly connected, signaling proteins themselves. When applied to early-stage design, the allo-network method is predicted to increase the likelihood that a given allosteric effector will proceed through the drug-approval process.[116, 122, 162]

Examples of recently discovered allosteric drugs

Several published examples for recently approved allosteric drugs serve to illustrate the current state-of-the art for emerging allosteric drug design principles. They also represent the significant advances that have been made to utilize structure-based methods for challenging druggable sites such as protein-protein interfaces and for membrane proteins such as ion channels. The available details of the discovery and optimization of these compounds do not include the methods discussed in this review, however they highlight where these predictive techniques could contribute to the allosteric design process.

In 2011, Gilmartin et al. of GlaxoSmithKline reported the discovery of a pharmacokinetically-optimized allosteric MEK inhibitor, GSK1120212.[163] The first generation inhibitor was discovered by high throughput screening [164] and the subsequent ternary crystal structure showed the allosteric pocket to be adjacent to the orthosteric ATP-bound site.[165] By 2012 there were fourteen allosteric MEK1/2 inhibitors in clinical trials,[166] because it was recognized that an inhibitor developed for this allosteric pocket afforded two very unique opportunities to avoid adverse clinical effects; the high doses required to compete against 1mM cellular ATP concentrations and inhibition of closely related ATP-binding sites in other kinases. The unique efficacy properties of GSK1120212 highlight both the opportunity and challenge of allosteric drug design. Gilmartin et al. report that although some other MEK allosteric drugs demonstrate inhibition of the ERK1/2 pathway *in vitro*, this has not translated into efficacy in patients.[163] GSK1120212 has since been approved in the U.S. under the name Trametinib for treatment of metastatic melanoma caused by the V600E mutation.

In 2012, Saalau-Bethell et al. of Astex reported the discovery of allosteric inhibitors for the HCV NS3 protein [126]. These inhibitors produce an allosteric effect by binding at an interdomain interface and stabilizing a pre-existing autoinhibited state of the protein. The original discovery of the allosteric site was accomplished using a fragment-based HTS technique, followed by optimization using X-ray crystallography and structure-based SAR. The authors discuss their experience with a few confounding factors in the allosteric design process, namely the need to use the full-length protein in their screening construct to observe the exerted allosteric effect, and the subsequent directed evolution study of resistance mutations that could occur at the allosteric site.

Hackos et al. of Genentech published on the discovery of positive allosteric modulators (PAMs) for GluN2A-containing NMDA receptors in 2016 [167]. PAMs are allosteric ligands which increase the effect of the endogenous signalling molecule and do not cause a change in its absence. The allosteric site in this case was at a protein-protein interface, and was discovered using HTS. Subsequent medicinal chemistry efforts then optimized the early hit molecule. The authors note that the validation of this allosteric site was reinforced by its similarity to an analogous allosteric site in AMPA receptors, but that the NMDA receptor site has elements of asymmetry that the AMPA receptor site did not. In comparing two similar compounds, GNE-

6901 and GNE-8324, the authors make comments that indicate evidence of mode switching or a shallow SAR landscape, and they further characterize the details of the allosteric mechanism using mutagenesis experiments.

In summary these examples demonstrate allosteric drug discovery can be successful at protein sites often considered to be undruggable. It is apparent that these successes can be further built upon through computational methods that allow for rational rather than serendipitous HTS discovery of new allosteric binding sites and a deeper understanding of allosteric mechanisms that overcome design challenges such as mode switching.

Protein-Sequence Analysis Methods

Introduction

Protein-sequence analysis is a useful tool to detect and characterize allosteric pathways and pockets. Here, we classify sequence-based methods into two groups: 1) "**single site"** methods, which produce a list of individual functional sequence positions; and 2) "**coupled site"** methods, which produce a list of groups comprised of two or more sequence positions that appear to be functionally linked based on their coevolution.

All sequence-based analysis methods share some challenges. These challenges include how to: select and aggregate clean, relevant sequences as input; interpret the output; and integrate sequence-analysis results with other forms of data. Determining the biological meaning of a strong signal is also problematic. While many analysis methods identify evolutionarily important residues, the specific biological role of these residues cannot be inferred without additional knowledge. For example, it is difficult to determine, based on sequence alone, whether an evolutionarily significant residue plays an allosteric role, or whether its role is related to another essential process (eg substrate binding, maintaining protein structure, etc.).[168, 169] Indeed, it is likely that a given residue serves multiple purposes simultaneously.

Input sequence selection and alignment also present challenges. Most techniques require many sequences to establish statistical significance. To obtain the required number of sequences, researchers often lower the stringency of their search parameters, resulting in alignments that contain sequences with lower similarity or incomplete coverage of the original query. While some analysis methods manage to detect meaningful coevolution over a wide range of multiple sequence alignment (MSA) conservation and noise levels, others are more susceptible to messy data.[170, 171] For a more complete discussion of these topics and how they affect coevolution analysis methods, readers are directed to an excellent recent review by Juan et al.[172]

Single-Site Evolutionary Analysis Methods

By our definition, single-site evolutionary analysis methods return a list of predicted functional sequence positions but do not suggest specific linkages between sites. Once a researcher has constructed an MSA, the conservation or phylogenetic relevance of each column can be used to infer the evolutionary importance of each sequence position. This importance is sometimes a hallmark of thermodynamically critical residues that participate in allostery. Though single-site methods only return a list of single high-scoring sequence positions, the inner workings of some single-site methods are based on the aggregate or correlated behaviors of multiple sequence positions (eg to determine baseline residue probabilities within a multiple-sequence alignment or construct a phylogenetic tree).

Single-site methods for detecting allostery are advantageous because they lack much of the noise often associated with correlation analysis. These analyses are also appealing because of their simplicity: There are usually fewer parameters to set, and the results can be visualized directly by highlighting key residues on a 3D protein structure.

Single-Position Entropy

Shannon entropy, one of the simplest nontrivial sequence-analysis metrics,[173] was used widely in early works to identify conserved sequence positions for drug-design or mutagenesis experiments.[174] Similar in form to thermodynamic entropy from statistical mechanics, Shannon entropy measures the population diversity of residues at a single MSA position. It is also central to mutual information (MI), a popular coupled-pair metric. The MI of two sequence positions is defined as the sum of the individual position entropies, minus the entropy of the positions considered jointly. While we not cover the mathematical details of these methods here, interested readers are directed to previous articles on these topics.[171, 175]

Shannon entropy does not consider amino acid similarity (eg in the Shannon-entropy framework, a leucine-to-isoleucine mutation is considered mathematically equivalent to a leucine-to-arginine mutation). Other entropy measures, such as the relative Shannon entropy (also called the Kullback-Leibler Divergence (KLD)[**176**] and the von Neumann entropy,[**5, 177**] attempt to overcome this limitation and, as a result, may be more useful in the search for allosteric sites. Relative Shannon entropy/KLD accounts for some measure of the protein's chemical environment by considering each mutation with respect to the background amino-acid frequencies calculated from the MSA. This analysis may be particularly useful when searching for allosteric sites in proteins that reside in membranes or other noncytosolic compartments, where background residue probabilities or mutational preferences may be biased due to different biochemical contexts. In contrast, von Neumann entropy, a concept borrowed from quantum statistical mechanics, is calculated using amino-acid similarity matrices. Identifying an optimal amino-acid similarity metric is nontrivial and may well depend on the nature of the system (eg in a well-packed protein, residue size may be a sensitive metric, whereas surface-site comparison may require the user to prioritize charge). In a recent publication describing these types of entropy, Johansson and Toh explored how the two metrics can be mixed to detect enzyme active sites with maximum sensitivity.[171]

Zhang et al. constructed a variety of new analysis methods in 2008 by combining Shannon or von Neumann entropy, phylogenetic tree structure, and a novel gap-treatment approach.[177] In benchmarking their method, they compared their results to Evolutionary Trace and ConSurf (discussed in greater detail below). Two of their hybrid approaches outperformed all other techniques in detecting significant residues across a variety of proteins: the Improved Zoom method, which incorporates a tree breakdown of subalignments, and the Physiochemical Similarity Zoom method, which extends the Improved Zoom method with von Neumann entropy and tree-branch-size normalization.

Evolutionary Trace.

Lichtarge, Bourne, and Cohen pioneered the evolutionary trace (ET) method. The approach has become quite popular, largely because the algorithm is intuitive and its results are readily visualizable.[178] ET aligns a number of sequences and constructs a phylogenetic tree, then monitors the conservation of sequence positions at major tree branching points. By slicing the tree at different similarity cutoffs, the algorithm extracts the cluster-defining sequence positions. The evolutionary significance of these sequence positions is implied by their conservation in the sequences beyond the next branch. In their first paper,[178] the authors demonstrated that ET can detect functionally important sites in SH2, SH3, and DNA-binding domains. Work has since been published on ET validation, parameter optimization, and best-use practices.[179]

In a method often referred to as "Difference-ET," the user runs ET on two related proteins and considers differences in the high-ranking residues and their scores. The sequence positions with strongly varying scores may suggest specificity determinants or differences in allosteric and/or orthosteric mechanisms. Notably, Difference-ET has been used in the study of GPCR specificity.[179-181]

To better account for varying rates of evolution in different subtrees and correlated mutations, in 2004 Mihalek et al. developed real-valued ET.[182] This method incorporates entropy information into the standard ET framework. This work also introduces the zoom ET method (not related to Improved Zoom, above), which adds higher weight to sequences that are more similar to the protein of interest. In the introductory work, they used real-valued and zoom ET to detect the functional residues in a kinase domain, then compared the performance of both methods to regular (integer-valued) ET and entropy. Given unpruned sequence data sets, the real-valued ET and zoom ET methods outperformed the others by a significant margin. In contrast, integer-valued ET prevailed in most respects when pruned data was available. An automated web server is available to perform real-valued ET calculations, generate reports, and visualize results at http://mammoth.bcm.tmc.edu/ETserver.html.[183]

H2r(s).

In 2008, Merkl et al. introduced a method called H2r that serves as a segue between single-site and coupled-site approaches.[184] H2r generates a mutual-information matrix for an MSA, then discards all but the strongest detected coupled pairs. For each sequence position k, the method returns conn(k), the number of top-ranked pairs that include k. Initial work proved that H2r can successfully detect functionally significant residues across a range of proteins. More recently, H2rs, an improved version of H2r, has been released.[185] This method modifies the original by using von Neumann instead of Shannon entropy and performing more detailed checks for statistical significance. H2rs is available as a web server and a stand-alone application at http://www-bioinf.uni-regensburg.de/.

Coupled-Site Evolutionary Analysis Methods

Second-order sequence analysis detects residue pairs that mutate in concert more frequently than would be expected given random genetic events. Coevolving residue pairs are assumed to be functionally linked, often because they serve essential roles in allostery or structural integrity.

The immediate output of second-order allostery analysis is a list of residue pairs with associated correlation strengths. Combining these individual pairwise correlations into a single picture of the entire protein is a separate task. On the most basic level, the strongest correlations that include a residue or site of interest can suggest thermodynamic coupling to other sites, possibly related to allostery. More complex analyses use hierarchical clustering or principal component analysis to analyze these linkages and uncover strongly linked networks of coevolving residues.

Basic Coupled-Site Analyses.

Several simple yet reliable residue-coupling analyses have maintained a presence in the literature over the past decades. These basic approaches are advantageous because they are easier to understand and have been shown to score consistently well in a wide range of tests. However, they may fail to detect correlations in more complex cases.[186, 187] Though more complex methods exist, many of these basic methods still appear as analysis options in coevolution-detection software packages and web servers. In this review, we focus on a few that are still widely used.

Mutual Information.

Mutual Information (MI) is one of the most straightforward and long-lived coupling metrics. The MI between two sequence positions is defined as the sum of the Shannon entropies of both positions, minus their joint entropy. Due to its simplicity and favorable mathematical properties, MI analysis is the basis for a number of more complex coevolution methods. However, MI does present certain shortcomings. For example, uncorrelated pairs of high-entropy sequence positions are likely to have a higher MI than uncorrelated pairs of low-entropy positions.[188, 189] To compensate for this and other shortcomings, various software packages have implemented a number of mathematical corrections to MI.[190-194] Further, methods to estimate baseline values for correlation (e.g., resampling or sequence shuffling) can improve MI analysis.[189, 195, 196]

Another relatively direct coevolution metric, the McLachlan-Based Substitution Correlation (McBASC),[197] looks for similar patterns of variation in the columns of an MSA, weighting for residue similarity using the McLachlan scoring matrix.[198] Analogous methods can be constructed using different substitution matrices, but McBASC continues to be a popular choice in the literature.[170, 189, 199]

In 2002, Kass and Horovitz[200] analyzed the GroEL complex using a chi-squared test to detect significant residue coevolution in an MSA. The analysis suggested intra- and inter-chain contact pairs and has continued to appear in the literature under the name “Observed Minus Expected Squared” (OMES).[170, 186, 187, 189, 201]

Statistical Coupling Analysis.

The Statistical Coupling Analysis (SCA) method developed by Lockless and Ranganathan is perhaps the most widely used sequence-based method for allostery prediction.[202] SCA draws an analogy to statistical physics by calculating a "coupling energy" between each sequence-position pair. The original SCA method computes a conservation value for each sequence position *i* in an MSA, applies one of several types of perturbation to another position *j* (depending on the SCA version,[203]) and finally recalculates the conservation at position *i* for the sequences that satisfy the perturbation. By calculating the change in individual and joint conservation over a variety of perturbations, SCA establishes a "coupling energy" that indicates the evolutionary coupling of positions i and j.

The output of the SCA method is an N x N matrix of coupling energies, where N is the number of sequence positions in the alignment. In early work, the researchers manually identified strongly coupled residue pairs that included one functional member (per experiment). More recent versions of SCA have grouped this matrix into meaningful clusters of coevolving residues using hierarchical or spectral clustering.[204]

Refinements of SCA have achieved improved statistical properties by resampling the original distribution.[205] In a 2011 paper, SCA was effectively used to engineer a light-sensitive LOV2 domain onto the surface of DHFR at a location that SCA had identified as energetically linked to the enzyme active site. Some variants of the resulting protein chimera were found to have acquired light-dependent activity.[206]

Further work has used SCA to design artificial WW domains.[207, 208] In 2011, an SCA analysis of antigen 85C from *Mycobacterium tuberculosis* suggested new sites that potentially could be exploited in drug design.[209] A number of projects have also demonstrated how SCA can be used to target mutations that affect protein function.[210-216]

Inspired by earlier work on the Sequence Correlation Entropy (SCE) method,[217] Dima and Thirumalai published an SCA variant in 2006.[218] This variant controls for specific protein composition by calculating the background probability that a given amino acid will be present at a random sequence position. This probability is determined by considering only the sequences being analyzed, as opposed to all sequences in the SWISS-PROT database.[219, 220] Further, they borrowed a coupled two-way clustering procedure from gene-sequence analysis to define the sectors.[221] In validating this method, the authors analyzed the PDZ, GPCR, and lectin families of proteins and were able to quantitatively predict functional residues, which were in agreement with experimental findings.

Explicit Likelihood of Subset Co-variation.

As mentioned above, SCA is a "perturbation-based" method in which correlation is established by excluding certain sequences from an MSA and monitoring how entropies change. Another popular perturbation-based method was published in 2003 by Dekker et al.[222] This method, Explicit Likelihood of Subset Co-variation (ELSC), relies on similar principles but returns correlation scores in the form of probabilities rather than statistical energies. ELSC was shown to be superior to SCA in contact prediction when tested on a range of protein families. It has since been implemented on web servers[223, 224] and has been a popular benchmark method in the literature.[186, 187, 199, 201]

Direct Coupling Analysis.

In 2009, Weigt et al. proposed a mutual-information-based method called Direct Coupling Analysis (DCA) that disentangles directly interacting residues from large networks of indirectly coupled sequence positions.[225] While this method is typically used in structure prediction to identify spatially adjacent sequence positions, it may find application in the study of short-range allosteric interactions. A more efficient implementation of the DCA method, known as "mean field" (as opposed to the original "message-passing" implementation) was published in 2011.[226] Both introductory papers show that DCA is a robust predictor of both intra- and inter-protein contacts and that it can hint at the existence of unobserved protein conformations. Related work has shown that DCA can be used in conjunction with structural models to generate predictive models of protein complexes,[227-229] determine the sequence positions that contribute to protein-interaction specificity,[230] and describe the conformational ensembles of proteins in crystallographic or near-crystallographic states.[231] A web server and software package are available to perform DCA analysis at http://dca.rice.edu/portal/dca/home.

PSICOV.

PSICOV is another popular contact-detection method that may find productive use in the study of allostery.[232] Mathematically, PSICOV relies on an estimated inverse of the MSA covariance matrix, which acts as a matrix of correlations between all sequence-position pairs that inherently controls for the variations in all other positions. PSICOV was successful at predicting contacting protein residues based on MSA data. The code has been published online at http://bioinfadmin.cs.ucl.ac.uk/downloads/PSICOV/.

Recurrence Quantification Analysis.

Recurrence Quantification Analysis (RQA), another second-order sequence-analysis technique, is best used when much is already known about the mechanism under investigation (eg physiochemical amino-acid properties such as charge or hydrophobicity are known to drive the allostery). RQA itself is a general method in nonlinear dynamics[233]: In the context of protein sequences, it considers a scalar-value vector that represents some property of a given sequence. In introductory work by Zbilut et al.,[234] the method was used to properly classify 56 TEM-1 beta-lactamase mutants with impaired function based on their hydrophobicity profiles. Further RQA work used hydrophobicity scores to classify proteins as allosteric or nonallosteric,[235] study p53 mutants,[236] and reveal interaction partners in viral-envelope proteins.[237]

In 2005, Colafranceschi et al. investigated the effect of choosing different physicochemical amino-acid descriptors and changing the numerical parameters of the RQA algorithm.[238] More recently, a comparison method based on RQA measurements, known as cross-RQA, effectively detected protein allostery.[239] Interested readers are directed to a review by Zbilut-Webber, which provides examples of RQA applied to a range of computational biology problems.[240]

Comparative Analyses.

Some work has been done to competitively benchmark the performance of these methods. In 2004, Fodor and Aldritch compared OMES, MI, SCA, and McBASC in a variety of tests. In short, the study found that performance is largely dependent on the way that different methods determine background residue probabilities and handle positional conservation.[170] A follow-up study investigated how effectively coevolution analysis finds thermodynamically linked residue pairs.[169] In general, spatially contiguous linked pairs were detected, but long-range couplings did not agree with experiments.

In 2010, Brown and Brown introduced a new pair-scoring method, called Z-scored-product Normalized Mutual Information (ZNMI), and compared it to the accuracy and reproducibility of MI, two versions of SCA, OMES, and ELSC.[187] The authors presented a thorough meta-analysis of method performance and the impact of input-parameter selection. Though none of the tools tested was particularly powerful, ZNMI was the most robust prediction tool. Brown and Brown also found that the use of multiple subalignments produced more accurate and reproducible results.

A comparative analysis of SCA and DCA revealed that the top 35 "sectons" found via spectral clustering of the DCA matrix corresponded to pairs, triplets, and quadruplets of spatially contiguous residues.[241] In contrast, a similar analysis of the SCA matrix produced spatially adjacent clusters of many residues each. These different results validate the stated goals of each method: DCA aims to find contacting pairs, whereas SCA aims to find potentially distant groups that are thermodynamically linked in a certain function.

In 2014, Pele et al. investigated seven coevolution analysis methods to find the hallmark covarying pairs in GPCR alignments.[186] They considered three variants of MI, McLachlan Based Substitution Correlation, SCA, ELSC, and OMES. OMES and ELSC were the most robust methods for finding the residues responsible for subfamily divergence. Their article also included an insightful discussion of the methods.

Mao et al. published a comparative analysis in 2015.[242] Their study tests OMES, two variants of MI, SCA, PSICOV, and DI, and finds that PSICOV and DI are best at identifying contacting residues. OMES and MIp excel at removing false positives from the lists of predicted contacts. While the authors focused on detecting inter- and intramolecular contacts, their analysis also provided useful insights to guide the productive use of each method. For example, all methods benefit from repeatedly shuffling the MSA and rerunning the analyses in order to provide a baseline and remove false positives. Finally, the authors found that the consensus of DI and PSICOV provides a more robust prediction of contacting residues than any single prediction method alone. The software used to perform this analysis is available through the ProDy Evol program http://prody.csb.pitt.edu/evol/}.[243]

In the course of introducing new types of MI analysis (dbZPX2, dgbZPX2, and nbZPX2) and evaluating the effectiveness of MSA simulation (a topic beyond the scope of this paper), Ackerman et al. in 2012 compared many different coevolution analyses in their ability to predict contacting residue pairs.[201] These comparisons found that the "new" methods (the ZPX2 family, DCA, and log(R) (not discussed here)), were significantly superior to the "old" methods (OMES, McBASC, ELSC, and SCA).

Web Servers.

Several web servers perform and visualize sequence analyses. Given a PDB code, Contact Map WebViewer (CMWeb)[224] automatically constructs an MSA and visualizes a variety of coevolution analyses: mutual information, SCA, ELSC, OMES, and an early method presented by Gobel et al.[244] The same server can also compare the results of these methods to user-uploaded data (eg results the user obtained using some other type of analysis). The CMWeb server can be accessed at http://cmweb.enzim.hu/.

The Coevolution Analysis of Protein Residues server hosted by the Gerstein Lab[223] http://coevolution.gersteinlab.org/coevolution/ can perform a large number of the coupled-site analyses presented in this review, including SCA, ELSC, MI, and McBASC-type methods employing different scoring matrices. The server can also validate the results of these methods in predicting residue distances in a crystal structure.

MISTIC (Mutual Information Server to Infer Coevolution) is an automated web server that accepts user-submitted MSAs or collects them from PFAM.[188] MISTIC uses a corrected form of MI to infer coevolving pairs and offers several analysis methods that combine structure and coevolution.[193] It can be accessed at http://mistic.leloir.org.ar/.

CAPS (Coevolution Analysis using Protein Sequences) is a unique algorithm that combines phylogenetic, 3D, and MSA data to predict coupled sequence positions.[245, 246] Versions 1 and 2 are hosted on web servers at http://bioinf.gen.tcd.ie/caps/ and http://caps.tcd.ie/, respectively.

The Interprotein-COrrelated Mutations Server (I-COMS, http://i-coms.leloir.org.ar/index.php} focuses on detecting contacts at protein-protein interfaces, though it can also return intra-chain correlations.[247] The server automatically builds alignments; performs MI, DCA, or PSICOV analysis; generates visualizations of the results; and allows users to download data taken from various points in the data-collection and analysis workflow.

In 2012, Jeong and Kim published a study describing a close MI variant.[248] They employed an automated workflow to control for various types of noise in sequence alignments, using the MSA sequence profile to establish prior knowledge about the protein. While the authors only studied a few MI variants, they stressed that their profile-based method could be extended to more complex analysis techniques. Their approach, Correlated Mutation Analysis Tool (CMAT), is available on a web server at http://binfolab12.kaist.ac.kr/cmat/.

Access to ConSurf, a single-site detection method similar to ET, is available at http://consurf.tau.ac.il/.[249-254]

Software.

The ProDy Python package, referred to above, can compute a variety of coevolution metrics.[242, 255, 256] In 2014, Skjaerven et al.[257] released the latest version of the powerful Bio3D R package for protein structure and sequence analysis.[258] While it focuses on structural analysis, the package can compute Shannon entropy and offers useful functions to create and manipulate sequence alignments. Also in 2014, Li et al.[259] published the CorMut software package for R, which computes MI, a metric called the "Jaccard index."[260] and the conditional selection pressure metric K_a/K_s.[261]

	Web server	Downloadable	Generates MSA	Shannon Entropy	Relative Shannon/KLD	MI	SCA	DCA	PSICOV	OMES	ELSC	Notes	URL
CMWeb[137]	X	X	X			X	X			X	X	Also computes an early method from Gobel et al.	http://cmweb.enzim.hu/
MISTIC[83]	X				X	X						Calculates a corrected version of MI	http://mistic.leloir.org.ar/
CMAT[141]	X	X	X			X						Many noise-filtering parameters to customize; returns MI, MIp, and MIc	http://binfolab12.kaist.ac.kr/cmat/
I-COMS[140]	X		X			X		X	X			Offers two variants of DCA	http://i-coms.leloir.org.ar/
ET[78]	X	X										Runs Evolutionary Trace	http://mammoth.bcm.tmc.edu/ETserver.html
Coevolution Analysis of Protein Residues[116]	X	X				X	X			X	X	Also computes similarity matrix-based methods (including McBASC), Chi Square, and Quartets coevolution metrics	http://coevolution.gersteinlab.org/
ConSurf[147]	X		X									Performs a single-site type of analysis similar to ET	http://consurf.tau.ac.il/
DCA[118,119]	X	X										Returns pairwise DI	http://dca.rice.edu/portal/dca/home
CAPS[138,139]	X	X										Runs a nonstandard coevolution analysis technique	http://caps.tcd.ie/
H2r[79]	X											Runs a nonstandard single-site coevolution analysis technique	http://www-bioinf.uni-regensburg.de/
H2rs[80]		X										Runs a nonstandard single-site coevolution analysis technique	http://www-bioinf.uni-regensburg.de/
Bio3D[150]		X		X								R package; useful for creating and modifying sequence alignments	http://thegrantlab.org/bio3d/
CorMut[152]		X				X						R package. Also computes Ka/Ks and Jaccard index	https://www.bioconductor.org/packages/release/bioc/html/CorMut.html
ProDy[148,149]		X		X		X	X			X		Python package; can compute DI and mutual information correction/normalization	http://prody.csb.pitt.edu/evol/

Figure 4.2: Selected coevolution web servers/software packages and their capabilities

Conclusions

Over the past decade, advances in computing power and predictive algorithms coupled with the increased availability of structural and biochemical data have revealed new opportunities for rational design of allosteric drugs. The emergence of novel computational approaches to describe and predict allosteric phenomena across a range of scales, from the coordinated atomic movements in a single receptor molecule to complex allosteric signaling networks, is ushering in a new era wherein computational methods can be used to prospectively predict, discover, and characterize allosteric sites and effector molecules. Within the context of a drug-discovery program, such approaches hold the potential for developing drugs with increased specificity and selectivity, as well as the ability to gain new and more comprehensive understanding of old targets. For example, the convergence of advances in (i) theoretical MSM-based frameworks and MSM building software, (ii) community MD codes that can achieve >100 ns/day sampling for realistic sized systems on single gaming/commodity GPU processors, and (iii) pocket and druggable site-detection algorithms now make it possible for researchers even in industrial settings, with fast-paced timelines and stringent quality standards, to apply these approaches to drug targets already in their arsenals. The application of these methods to kinases and GPCRs seems particularly worthwhile, given the existence of assays and structural data, and the challenges faced by existing drug candidates in the clinic.

Acknowledgements

Chapter 4 is a modified, partial reprint of the material as it appears in "Wagner, J.R., Lee, C.T., Durrant, J.D., Malmstrom, R.D., Feher, V.A., Amaro, R.E. (2016) Emerging Computational Methods for the Rational Discovery of Allosteric Drugs, *Chemical Reviews* doi:

10.1021/acs.chemrev.5b00631”. The dissertation author was one of two primary authors of the paper.

This work was funded in part by the Director's New Innovator Award Program (NIH DP2 OD007237 to REA) and the National Science Foundation through XSEDE supercomputer resources provided via TG-CHE060073N to REA. Funding and support from the National Biomedical Computation Resource was provided through NIH P41 GM103426. JRW and CTL are supported by a NIH Molecular Biophysics Training Grant T32 GM008326. Additional resources from the UC San Diego Center for Drug Discovery Innovation and the Keck II Computing and Visualization facility are gratefully acknowledged.

Chapter 5 : Continuous Evaluation of Ligand Pose Prediction (CELPP) Challenge: A Tool to Evaluate and Improve Protein-Ligand Docking Methods

Introduction

Determining a therapeutic ligand's pose inside of a protein binding site can greatly accelerate rational drug design efforts[4, 262]. While it is possible to determine a bound ligand pose using experimental structure-determination methods like X-ray crystallography, the process is difficult and time-consuming. To accelerate drug discovery efforts, "docking" algorithms have been developed, which predict how ligands bind to proteins[6, 263-267]. These algorithms have enjoyed years of public and commercial development, both in the form of optimization and invention of novel approaches. Currently, many drug design efforts take advantage of docking algorithms. However, there is a lack of standardization, or well-understood "best practices" for how to most effectively select and use algorithms for specific types of problems[268-271].

Previous efforts to benchmark biomolecular structure prediction algorithms have been well received. These efforts succeeded by achieving a high throughput of test cases, formalizing common approaches using automated workflows, and engaging with the scientific community to define goals and disseminate results[272]. Notable previous efforts include CSAR[273-277], CASP[278-280], GPCRDOCK[281-283], and CAMEO[284].

The Drug Design Data Resource (D3R; drugdesigndata.org) is a NIH funded resource aimed at providing benchmark datasets and blinded challenges to assist in the evaluation and improvement of computer-aided drug design (CADD) algorithms[16, 285-287]. Previous blinded challenges hosted by D3R have enjoyed broad community participation, but are difficult to scale up due to their reliance on donations of structural datasets from private groups and the necessary curation of donated data. The results of these competitions are often hard to draw conclusions from, as the leading participants in D3R Grand Challenges report using a variety of algorithms and diverse strategies [286, 287]. While the D3R Grand Challenge activities will continue, a new challenge format has been sought that more rigorously documents the methods used for predictions and increases the number of predictions made.

To further the mission of D3R, we introduce Continuous Evaluation of Ligand Pose Prediction (CELPP), a rolling, weekly challenge for automated pose predicting tools. CELPP is based on the Protein Data Bank (PDB) weekly release of forthcoming structures, and may be summarized as follows. By processing the weekly Protein Data Bank report of upcoming structures, we identify ~40 soon-to-be-released protein ligand complexes as “targets”, that are suitable for benchmarking pose prediction algorithms. For each target, D3R suggests docking to homologous, already-released “candidate” structures that are suitable for cross-docking. Files containing information about the targets and the candidate structures are sent to CELPP participants, who have a set amount of time to predict how the ligand binds. The participants submit their predictions to a D3R server before the release of the new protein-ligand crystal structures, and D3R evaluates the correctness of each prediction. These performance statistics will be visualized and published for further analysis.

CELPP challenge participants implement their computational docking workflow on their own server. To lower the barrier to challenge entry, D3R provides CELPPade. CELPPade is a Python framework for participant servers that receives the weekly D3R challenge package, applies the contestant’s prediction workflow, and uploads the predictions back to D3R for evaluation.

It is useful to define two separate concepts in this work -- “pose prediction” and “affinity prediction”[265]. CELPP focuses on pose prediction -- that is, taking a ligand which is known to bind to a protein, and computationally predicting its 3D placement and conformation in the binding pocket. CELPP does not focus on affinity prediction, which is defined as the process of determining how strongly a given ligand binds to a protein. Though conceptually overlapping, we do not assume that an algorithm which is ideal for one task is also well suited for the other. Each task is performed under different resource constraints. For example, an algorithm which accurately predicts binding poses but requires one day to run is practical for most pose prediction cases. However, the same algorithm would not be feasible for screening hundreds of thousands of ligands in a screening campaign, and would not necessarily be able to compare different molecules on the basis of affinity. For this and other reasons, we treat pose prediction as an independent problem from affinity prediction.

Further, it is useful to discuss the “scope” of CELPP based on how problems are encountered in real-world applications. CADD scientists are likely to use pose prediction algorithms after a ligand is experimentally found to bind to a protein of interest. The ligand at that point is known by its 2D structure, and in many cases the protein (especially those of known therapeutic interest) will have had its 3D structure solved and deposited in the Protein Data Bank. These 3D structures may be bound to a similar ligand, a dissimilar ligand, or no ligand at all. Using these pre-existing structures to predict the binding pose of a new ligand is referred to as “cross docking”[265, 288, 289] To account for these cases, the candidate structures selected by D3R for the CELPP challenge are selected to represent diverse amounts of prior information.

This challenge scope defines a set of common tasks that CELPP participants must perform. Protein structures from the Protein Data Bank often require processing before they can be used for pose prediction[290, 291]. This processing can be in the form of resolving ambiguously assigned electron density, removing or retaining solvent molecules, accounting for crystallization artifacts, and numerous other areas. Similarly, the process of using a 2D ligand structure to ultimately determine a 3D binding pose requires exploring many possible conformations and protonation states, among other issues[28, 290, 292]. In CELPP, we refer to these steps as "preparation", and encourage participants to explore and optimize different approaches to these problems. After appropriate molecule preparation, a docking algorithm can be run to predict the ligand pose inside of the binding site.

Ultimately, the goal of CELPP is to accelerate the development of CADD algorithms by identifying strengths and shortcomings of modern techniques. Due to the large number of targets per week, we expect to achieve a greater level of statistical significance than the D3R Grand Challenges. Further, it will be possible to identify preparation and docking algorithms that are best-suited to different classes of problems, such as cytosolic vs. membrane proteins, or flexible vs. rigid ligand docking[267]. Given a large number of targets and participants, we hope also to deconvolute the contributions of each step in these approaches.

Results

D3R-side implementation

Target and candidate selection

D3R hosts the weekly CELPP challenge by scanning the Protein Data Bank (PDB)[57, 293, 294] pre-release announcements and applying filters to identify a subset of the entries as **targets** for the challenge. Targets are 3D protein-ligand structures which are in the final stages of processing at the PDB but do not yet have their 3D coordinates released. At the time of the challenge the only public data available about a target is its sequence and the 2D structure of the bound ligand. For each target, up to 5 already-released structures of the same protein, called **candidates**, are identified from the PDB as being suitable for cross-docking (Figure 5.1). This process mimics a popular strategy that participants in previous D3R challenges used to select structures for pose prediction.

Suitable **target** sequences:

- Have only one unique protein sequence (to exclude hetero-oligomers)
- Have only one ligand (excluding metals, solvents, and other non-druglike molecules)
- Has at least one **candidate** structure (defined below)

Suitable **candidate** structures for a target:

- Have >95% sequence identity with the target sequence
- Have >90% sequence coverage with the target sequence
- Have only one unique protein sequence (to exclude hetero-oligomers)
- Are determined via X-ray crystallography

To simulate a variety of realistic pose prediction scenarios, up to five structures are selected from the set of suitable candidates for the CELPP challenge, as follows:

- **LMCSS:** The candidate structure with the **largest maximal common substructure** to the target ligand is selected. The center of mass of the ligand in this structure is used to suggest the binding pocket for all predictions. In the case that two candidate structures tie for the largest maximal common substructure, the highest-resolution candidate is used.

- **SMCSS:** The candidate structure with the **smallest maximal common substructure** to the target ligand is selected. In the case that two candidate structures tie for the smallest maximal common substructure, the highest-resolution candidate is used.
- **hiTanimoto:** The candidate structure with the **highest ligand Tanimoto** score to the query ligand is selected. In the case that two candidate structures tie for the smallest maximal common substructure, the highest-resolution candidate is used.
- **hiResHolo:** The candidate structure with the **highest crystallographic resolution and any druglike ligand** is selected.
- **hiResApo:** The candidate structure with the **highest crystallographic resolution and no druglike ligand** is selected

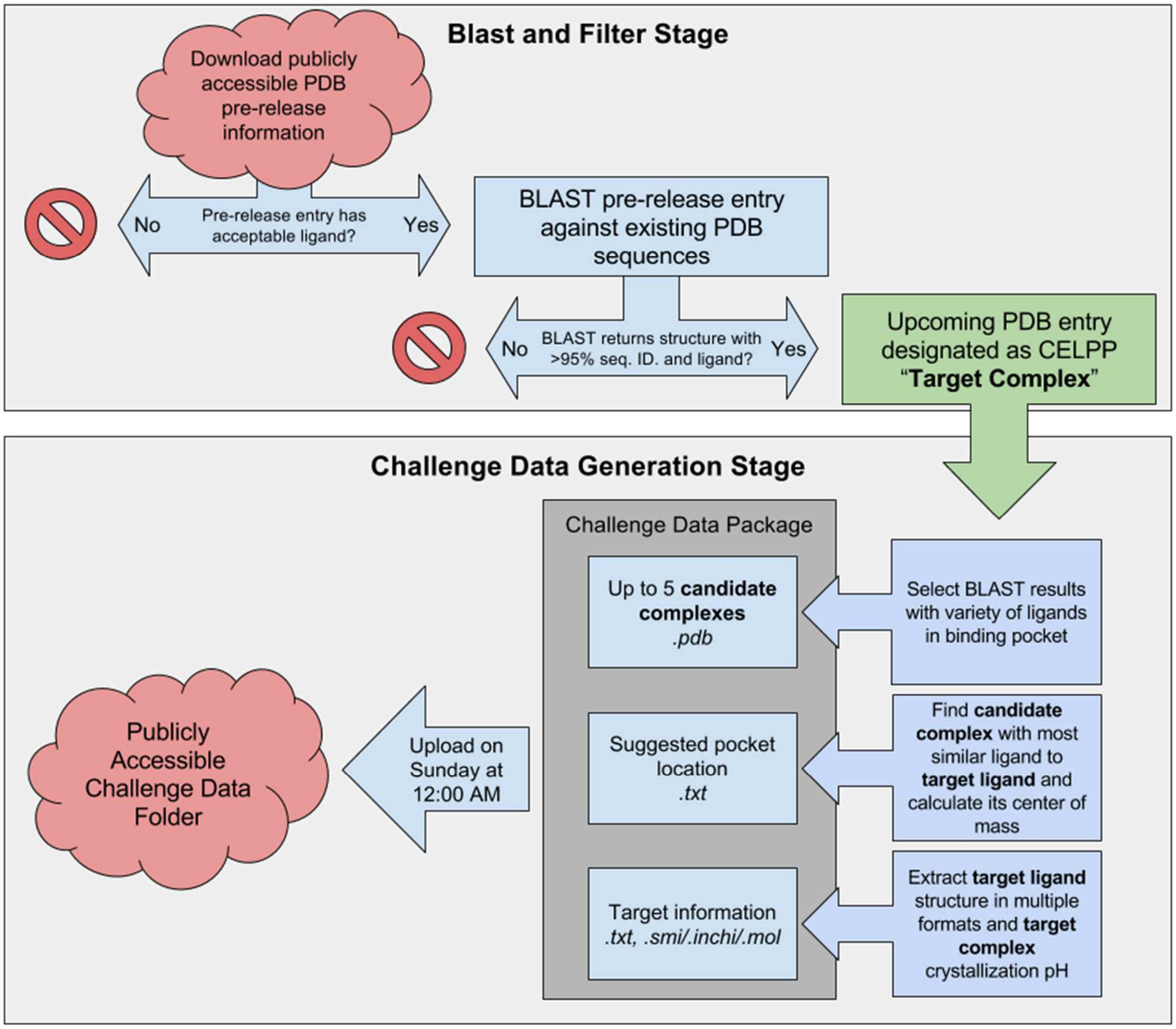

Blast and Filter Stage
Download publicly accessible PDB pre-release information
No
Pre-release entry has acceptable ligand?
Yes
BLAST pre-release entry against existing PDB sequences
No
BLAST returns structure with >95% seq. ID. and ligand?
Yes
Upcoming PDB entry designated as CELPP "Target Complex"
Challenge Data Generation Stage
Challenge Data Package
Up to 5 candidate complexes .pdb
Suggested pocket location .txt
Target information .txt, .smi/.inchi/.mol
Select BLAST results with variety of ligands in binding pocket
Find candidate complex with most similar ligand to target ligand and calculate its center of mass
Extract target ligand structure in multiple formats and target complex crystallization pH
Upload on Sunday at 12:00 AM
Publicly Accessible Challenge Data Folder

Figure 5.1 : CELPP Challenge Package Generation Scheme.

CELPP downloads the publicly-available PDB pre-release information, and then processes the new entries to assemble the weekly challenge package.

Challenge data package

The final set of valid targets and their respective candidates comprise the CELPP weekly "challenge data" package. The challenge data package is uploaded to a public Box.com folder roughly each Sunday morning around midnight U.S. Pacific time. For each target, the challenge data package contains:

- A text file containing relevant information about the ligand, crystallization conditions, and selected candidate structures
- SMILES, InChI, and 2D MOL files of the target ligand
- PDB structures of the candidate proteins, pre-aligned to a reference
- PDB structure of the LMCSS ligand
- The suggested binding pocket center (center of mass of the LMCSS ligand)

D3R provides a script, **getchallengedata.py**, that downloads the active CELPP week's challenge data package.

Prediction submissions

CELPP participants upload their pose prediction results to a private submission folder provided by D3R. This upload must be completed before 3 PM U.S. Pacific time on Tuesday to be considered valid for scoring.

For each candidate, a valid submission consists of:

- A receptor structure, in PDB format
- A ligand structure, in MOL format

Strict adherence to these file formats is required, and any deviation from the official format may result in improper scoring and disqualification of the submitted prediction.

D3R provides a script, **packdockingresults.py**, that accepts a formatted directory of docking results, and then compresses it into a tar file and uploads it to a participant's private submission folder.

Evaluation of predictions

Evaluation of submissions begins after the close of the submission window. Due to potential complexities of crystallographic data, such as ambiguous ligand chain assignments and multiple monomers with different ligand poses in the asymmetric unit, automated evaluation of docked poses is somewhat involved. A prediction is given the lowest RMSD that can be achieved by aligning to each crystal chain. Each prediction is evaluated according to Scheme 1.

```
RMSD_list = []
For crystal_chain in crystal_structure:
    predicted_complex = merge(predicted_chain, predicted_ligand)
    Align predicted_complex to crystal_chain
    aligned_predicted_protein, aligned_predicted_ligand =
split(predicted_complex)
    For heavy_atom_mapping in atom_symmetries(crystal_ligand,
predicted_ligand):
        RMSD_list.append(this_mapping_RMSD)
Return min(RMSD_list)
```

Scheme 1: Ligand pose evaluation

Currently, only RMSD is evaluated. However, other metrics such as protein-ligand interaction fingerprint, internal RMSD, and Real Space Correlation Coefficient (RSCC) may be implemented in the future.

Target labels

Targets in CELPP can present many complicating factors, such as closed binding sites or the inclusion of structural waters. While D3R does not know about complicating factors when the challenge package is generated, it is possible to identify these cases when the structures are released to the public at the end of the week. As some prediction methods will be able to correctly solve these challenging cases, it is important that these cases are identified when reporting performance. For that reason D3R plans to “label” these more challenging targets and put them in separate scoring pools. These labels are expected to include: ligands bound at homodimer interfaces, ligands bound in cryptic pockets, ligands which interact with cofactors, and ligands bound at a location other than where D3R suggested.

Participating in CELPP

Enrollment

D3R has linked registration instructions for CELPP on the main website (drugdesigndata.org). Registration will provide participants with upload/download credentials for CELPP submissions.

Prediction schedule

Participants in CELPP should make a workflow that is able to process up to 100 targets in the 63-hour submission window. Figure 5.2 shows the standard weekly schedule for CELPP. This requirement means that the workflow should be able to process 100 ligand preparation tasks, 400-500 protein preparation tasks, and 400-500 docking tasks in 63 hours. Participants in CELPP may choose to submit results for only a subset of the targets. There are no restrictions on hardware or parallelization, however D3R requests that workflows should run independent of human intervention. In the future, D3R may request that participants submit workflows in the form of scripts or machine images.

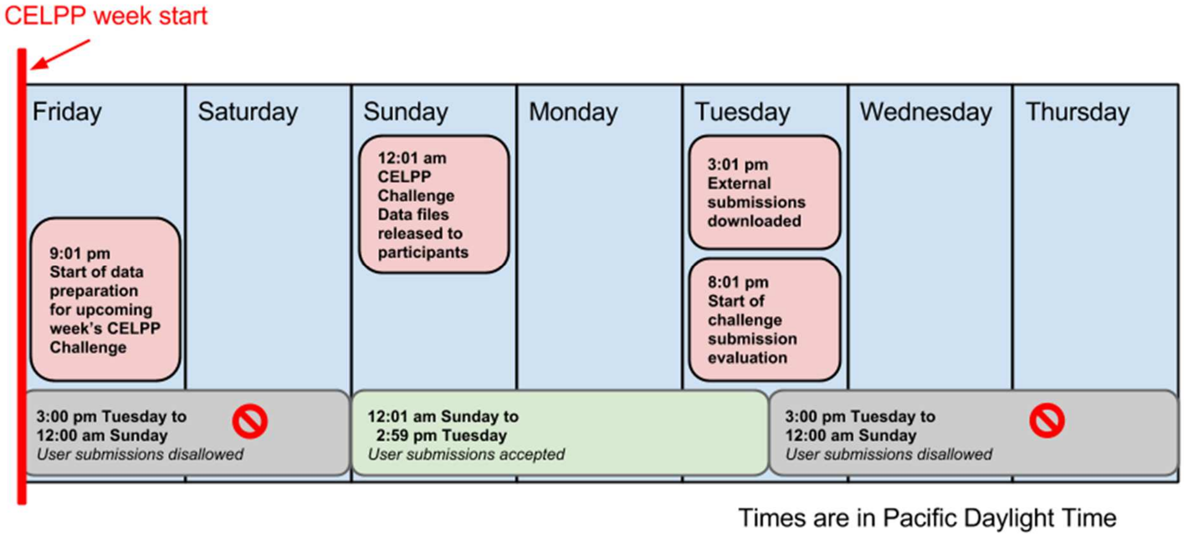

CELPP week start
Friday
Saturday
Sunday
Monday
Tuesday
Wednesday
Thursday
9:01 pm Start of data preparation for upcoming week's CELPP Challenge
12:01 am CELPP Challenge Data files released to participants
3:01 pm External submissions downloaded
8:01 pm Start of challenge submission evaluation
3:00 pm Tuesday to 12:00 am Sunday User submissions disallowed
12:01 am Sunday to 2:59 pm Tuesday User submissions accepted
3:00 pm Tuesday to 12:00 am Sunday User submissions disallowed
Times are in Pacific Daylight Time

Figure 5.2 : The CELPP week.

The CELPP week begins with the publication of PDB pre-release data on Friday evening. Challenge data preparation runs Friday evening and Saturday, and the upcoming week's challenge package is uploaded by the beginning of Sunday. Submissions are then accepted until Tuesday at 3:01 pm. Evaluation of the predictions begins on Tuesday evening.

Prediction workflows

To reduce the burden on participants, D3R provides two forms of assistance in creating pose prediction workflows.

The first form of assistance is the "CELPPade" Python package. CELPPade is a workflow template which contains empty Python functions that iterate over a challengedata package to perform user-specified protein preparation, ligand preparation, and docking. This blank workflow enables participants to run Python or shell commands, and each step is a Python function with set input and output file names. Figure 5.3 shows the files and functions that are exposed by CELPPade to enable creation of a modular pose prediction workflow.

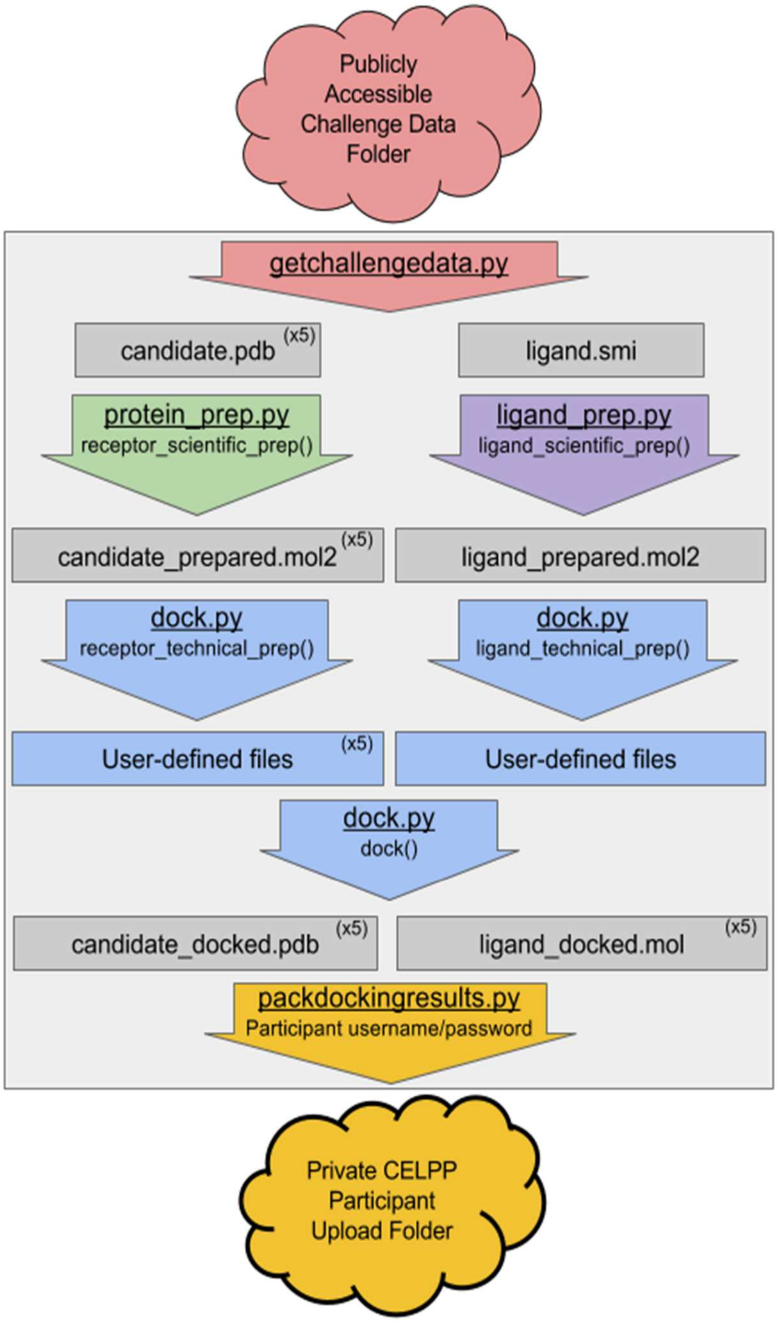


Figure 5.3 : Customization options of CELPPade workflow template.

Vertical arrows indicate functions, rectangles indicate files passed between workflow steps, and clouds represent internet-accessible folders. The large grey box indicates the steps that are run on the participant's computer. Different colors indicate script files for different steps of pose prediction.

The second form of assistance is a functioning workflow that runs Chimera DockPrep on both the protein and ligand, and then AutoDock Vina for pose prediction. This workflow is built using the CELPPade template and provides examples of running shell commands within Python. Care was taken to ensure that this workflow uses code which is free for use by academic labs, and can run on any computer with Python, Chimera, RDKit, OpenBabel, and Autodock Vina installed. Download and installation instructions for this workflow are provided on the CELPP website.

Participants are not required to use the CELPPade template -- it is provided as a convenience. Regardless of whether participants make use of the D3R-provided CELPPade template, the getchallengedata.py and packdockingresults.py scripts can be used in a standalone fashion to perform the weekly data download and upload. If D3R modifies the mechanism of providing challenge data packages or receiving participant predictions, new versions of these scripts will be made available.

Score reporting

Scores are emailed directly to participants. In the future scores will be posted online. Participants may choose to remain anonymous, in which case their results may be posted without identifying information.

Discussion

One goal of D3R is to encourage the creation of automated drug discovery workflows. The CELPP challenge data format is designed to be general in the information that it provides, rather than being shaped to the inputs of any specific docking program. Therefore, we anticipate that a workflow that is compatible with the CELPP challenge data format will be suitable for real-world pose prediction applications beyond benchmarking.

CELPP workflows are encouraged to perform protein preparation, ligand preparation, and docking in a modular fashion, where each is a standalone process. The benefit of using modular workflows is that researchers will be able to easily swap and study how each part influences workflow performance. This process should enable both direct comparison of distinct approaches, as well as benchmarking of incremental improvements.

D3R has established the Continuous Evaluation of Ligand Pose Prediction (CELPP) challenge to evaluate pose prediction workflow performance. This challenge offers benefits for both participants and the community by establishing a neutral benchmark set of targets and encouraging the automation of pose prediction workflows.

The rigorous standardization encouraged by CELPP will enable the measurement of how developments that are removed from the specific process of docking, such as structural water prediction and unresolved atom replacement, contribute to overall pose prediction accuracy. D3R recommends that steps in workflows be made modular to provide an easy means of swapping one process for another. The ability to A/B test the performance of steps in a docking workflow over large test sets will aid the identification of the best performing individual steps and combinations.

CELPP will identify the greatest strengths of current approaches, help map docking problems to the algorithms most likely to solve them, and illuminate areas of unmet need in structure-guided drug design. D3R will, via the analysis and publication of results, align community efforts to push forward the cutting edge of computer-aided drug design.

Acknowledgements

This work was done in conjunction with Shuai Liu, Robert V. Swift, Christopher P. Churas, Victoria A. Feher, Rommie E Amaro, and Michael K. Gilson. This work was supported by National Institutes of Health (NIH) grant 1U01GM111528 for the Drug Design Data Resource (D3R). RCSB Protein Data Bank is supported by NSF, NIH, and DOE (NSF DBI-1338415).

References

1. Durrant, J. D.; McCammon, J. A., Molecular dynamics simulations and drug discovery. *BMC Biol.* **2011,** *9* (1), 71.

2. Halgren, T. A., Identifying and Characterizing Binding Sites and Assessing Druggability. *J. Chem. Inf. Model.* **2009,** *49* (2), 377--389.

3. Jorgensen, W. L., The Many Roles of Computation in Drug Discovery. *Science (80-.).* **2004,** *303* (5665), 1813--1818.

4. Kuhn, B.; Guba, W.; Hert, J.; Banner, D.; Bissantz, C.; Ceccarelli, S.; Haap, W.; Körner, M.; Kuglstatter, A.; Lerner, C.; Mattei, P.; Neidhart, W.; Pinard, E.; Rudolph, M. G.; Schulz-Gasch, T.; Woltering, T.; Stahl, M., A Real-World Perspective on Molecular Design. *J. Med. Chem.* **2016,** *59* (9), 4087-4102.

5. Nussinov, R.; Tsai, C.-J., Allostery in Disease and in Drug Discovery. *Cell* **2013,** *153* (2), 293--305.

6. Sinko, W.; Lindert, S.; McCammon, J. A., Accounting for receptor flexibility and enhanced sampling methods in computer-aided drug design. *Chem. Biol. Drug Des.* **2013,** *81* (1), 41-49.

7. Betz, R. M.; Walker, R. C., Paramfit: automated optimization of force field parameters for molecular dynamics simulations. *J. Comput. Chem.* **2015,** *36* (2), 79-87.

8. Guvench, O.; MacKerell, A. D., Jr., Comparison of protein force fields for molecular dynamics simulations. *Methods Mol Biol* **2008,** *443*, 63-88.

9. Harder, E.; Damm, W.; Maple, J.; Wu, C.; Reboul, M.; Xiang, J. Y.; Wang, L.; Lupyan, D.; Dahlgren, M. K.; Knight, J. L.; Kaus, J. W.; Cerutti, D. S.; Krilov, G.; Jorgensen, W. L.; Abel, R.; Friesner, R. A., OPLS3: A Force Field Providing Broad Coverage of Drug-like Small Molecules and Proteins. *J. Chem. Theory Comput.* **2016,** *12* (1), 281-296.

10. Hornak, V.; Abel, R.; Okur, A.; Strockbine, B.; Roitberg, A.; Simmerling, C., Comparison of multiple Amber force fields and development of improved protein backbone parameters. *Proteins* **2006,** *65* (3), 712-725.

11. Jo, S.; Cheng, X.; Lee, J.; Kim, S.; Park, S. J.; Patel, D. S.; Beaven, A. H.; Lee, K. I.; Rui, H.; Park, S.; Lee, H. S.; Roux, B.; MacKerell, A. D., Jr.; Klauda, J. B.; Qi, Y.; Im, W., CHARMM-GUI 10 years for biomolecular modeling and simulation. *J Comput Chem* **2017,** *38* (15), 1114-1124.

12. Malde, A. K.; Zuo, L.; Breeze, M.; Stroet, M.; Poger, D.; Nair, P. C.; Oostenbrink, C.; Mark, A. E., An Automated Force Field Topology Builder (ATB) and Repository: Version 1.0. *J. Chem. Theory Comput.* **2011,** *7* (12), 4026-4037.

13. Vanommeslaeghe, K.; Hatcher, E.; Acharya, C.; Kundu, S.; Zhong, S.; Shim, J.; Darian, E.; Guvench, O.; Lopes, P.; Vorobyov, I.; Mackerell, A. D., CHARMM general force field: A force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields. *J. Comput. Chem.* **2009**, NA-NA.

14. Vanommeslaeghe, K.; Yang, M.; MacKerell, A. D., Jr., Robustness in the fitting of Molecular Mechanics parameters. *J. Comput. Chem.* **2015,** *36* (14), 1083.

15. 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.* **2004,** *25* (9), 1157-1174.

16. Yin, J.; Henriksen, N. M.; Slochower, D. R.; Shirts, M. R.; Chiu, M. W.; Mobley, D. L.; Gilson, M. K., Overview of the SAMPL5 host-guest challenge: Are we doing better? *J. Comput. Aided Mol. Des.* **2017,** *31* (1), 1-19.

17. Perez, A.; Morrone, J. A.; Simmerling, C.; Dill, K. A., Advances in free-energy-based simulations of protein folding and ligand binding. *Curr Opin Struct Biol* **2016,** *36*, 25-31.

18. Vanommeslaeghe, K.; MacKerell, A. D., Jr., CHARMM additive and polarizable force fields for biophysics and computer-aided drug design. *Biochim Biophys Acta* **2015,** *1850* (5), 861-71.

19. Indu Kumari, P. S., Mushtaq Ahmed and Yusuf Akhter*, Molecular Dynamics Simulations, Challenges and Opportunities: A Biologist's Prospective. *Current Protein & Peptide Science* **2018,** *18* (11), 1163-1179.

20. Genheden, S.; Ryde, U., The MM/PBSA and MM/GBSA methods to estimate ligand-binding affinities. *Expert Opin Drug Discov* **2015,** *10* (5), 449-61.

21. Durrant, J. D.; de Oliveira, C. A. F.; McCammon, J. A., POVME: an algorithm for measuring binding-pocket volumes. *J. Mol. Graph. Model.* **2011,** *29* (5), 773-776.

22. 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.* **2014,** *10* (11), 5047-5056.

23. Wagner, J. R.; Sørensen, J.; Hensley, N.; Wong, C.; Zhu, C.; Perison, T.; Amaro, R. E., POVME 3.0: Software for Mapping Binding Pocket Flexibility. *J. Chem. Theory Comput.* **2017,** *13* (9), 4584-4592.

24. Wagner, J.; Lee, C. T.; Durrant, J. D.; Malmstrom, R. D.; Feher, V. A.; Amaro, R. E., Emerging Computational Methods for the Rational Discovery of Allosteric Drugs. *Chem Rev* **2016,** *116* (11), 6370-90.

25. Brooijmans, N.; Kuntz, I. D., Molecular Recognition and Docking Algorithms. *Annu. Rev. Biophys. Biomol. Struct.* **2003,** *32* (1), 335-373.

26. Antunes, D. A.; Devaurs, D.; Kavraki, L. E., Understanding the challenges of protein flexibility in drug design. *Expert Opin. Drug Discov.* **2015,** *10* (12), 1301-1313.

27. Wang, J.; Wang, W.; Kollman, P. A.; Case, D. A., Automatic atom type and bond type perception in molecular mechanical calculations. *J. Mol. Graph. Model.* **2006,** *25* (2), 247-260.

28. Ebejer, J.-P.; Morris, G. M.; Deane, C. M., Freely available conformer generation methods: how good are they? *J. Chem. Inf. Model.* **2012,** *52* (5), 1146-1158.

29. Hawkins, P. C. D.; Skillman, A. G.; Warren, G. L.; Ellingson, B. A.; Stahl, M. T., Conformer generation with OMEGA: algorithm and validation using high quality structures from the Protein Databank and Cambridge Structural Database. *J. Chem. Inf. Model.* **2010,** *50* (4), 572-584.

30. Lexa, K. W.; Carlson, H. A., Protein flexibility in docking and surface mapping. *Q. Rev. Biophys.* **2012,** *45* (3), 301-343.

31. Lin, J.-H., Accommodating protein flexibility for structure-based drug design. *Curr. Top. Med. Chem.* **2011,** *11* (2), 171-178.

32. Spyrakis, F.; BidonChanal, A.; Barril, X.; Luque, F. J., Protein flexibility and ligand recognition: challenges for molecular modeling. *Curr. Top. Med. Chem.* **2011,** *11* (2), 192-210.

33. Ellingson, S. R.; Miao, Y.; Baudry, J.; Smith, J. C., Multi-conformer ensemble docking to difficult protein targets. *J. Phys. Chem. B* **2015,** *119* (3), 1026-1034.

34. Sørensen, J.; Demir, Ö.; Swift, R. V.; Feher, V. A.; Amaro, R. E., Molecular Docking to Flexible Targets. In *Methods in Molecular Biology*, 2014; pp 445-469.

35. Huang, S.-Y.; Zou, X., Ensemble docking of multiple protein structures: considering protein structural variations in molecular docking. *Proteins* **2007,** *66* (2), 399-421.

36. Totrov, M.; Abagyan, R., Flexible ligand docking to multiple receptor conformations: a practical alternative. *Curr. Opin. Struct. Biol.* **2008,** *18* (2), 178-184.

37. Wong, C. F., Flexible receptor docking for drug discovery. *Expert Opin. Drug Discov.* **2015,** *10* (11), 1189-1200.

38. Osguthorpe, D. J.; Sherman, W.; Hagler, A. T., Exploring protein flexibility: incorporating structural ensembles from crystal structures and simulation into virtual screening protocols. *J. Phys. Chem. B* **2012,** *116* (23), 6952-6959.

39. Henrich, S.; Salo-Ahen, O. M. H.; Huang, B.; Rippmann, F. F.; Cruciani, G.; Wade, R. C., Computational approaches to identifying and characterizing protein binding sites for ligand design. *J. Mol. Recognit.* **2010,** *23* (2), 209-219.

40. Kokh, D. B.; Czodrowski, P.; Rippmann, F.; Wade, R. C., Perturbation Approaches for Exploring Protein Binding Site Flexibility to Predict Transient Binding Pockets. *J. Chem. Theory Comput.* **2016,** *12* (8), 4100-4113.

41. Stank, A.; Kokh, D. B.; Fuller, J. C.; Wade, R. C., Protein Binding Pocket Dynamics. *Acc. Chem. Res.* **2016,** *49* (5), 809-815.

42. Kokh, D. B.; Richter, S.; Henrich, S.; Czodrowski, P.; Rippmann, F.; Wade, R. C., TRAPP: a tool for analysis of transient binding pockets in proteins. *J. Chem. Inf. Model.* **2013,** *53* (5), 1235-1252.

43. Kleywegt, G. J.; Jones, T. A., Detection, delineation, measurement and display of cavities in macromolecular structures. *Acta Crystallogr. D Biol. Crystallogr.* **1994,** *50* (Pt 2), 178-185.

44. Desaphy, J.; Azdimousa, K.; Kellenberger, E.; Rognan, D., Comparison and Druggability Prediction of Protein–Ligand Binding Sites from Pharmacophore-Annotated Cavity Shapes. *J. Chem. Inf. Model.* **2012,** *52* (8), 2287-2299.

45. Paramo, T.; East, A.; Garzón, D.; Ulmschneider, M. B.; Bond, P. J., Efficient Characterization of Protein Cavities within Molecular Simulation Trajectories: trj_cavity. *J. Chem. Theory Comput.* **2014,** *10* (5), 2151-2164.

46. Craig, I. R.; Pfleger, C.; Gohlke, H.; Essex, J. W.; Spiegel, K., Pocket-space maps to identify novel binding-site conformations in proteins. *J. Chem. Inf. Model.* **2011,** *51* (10), 2666-2679.

47. Liang, J.; Edelsbrunner, H.; Woodward, C., Anatomy of protein pockets and cavities: measurement of binding site geometry and implications for ligand design. *Protein Sci.* **1998,** *7* (9), 1884-1897.

48. Saberi Fathi, S.; Fathi, S. S.; Tuszynski, J. A., A simple method for finding a protein's ligand-binding pockets. *BMC Struct. Biol.* **2014,** *14* (1), 18.

49. Le Guilloux, V.; Schmidtke, P.; Tuffery, P., Fpocket: an open source platform for ligand pocket detection. *BMC Bioinformatics* **2009,** *10*, 168.

50. Schmidtke, P.; Le Guilloux, V.; Maupetit, J.; Tufféry, P., fpocket: online tools for protein ensemble pocket detection and tracking. *Nucleic Acids Res.* **2010,** *38* (Web Server issue), W582-9.

51. Laurent, B.; Chavent, M.; Cragnolini, T.; Dahl, A. C. E.; Pasquali, S.; Derreumaux, P.; Sansom, M. S. P.; Baaden, M., Epock: rapid analysis of protein pocket dynamics. *Bioinformatics* **2015,** *31* (9), 1478-1480.

52. Durrant, J. D.; McCammon, J. A., BINANA: a novel algorithm for ligand-binding characterization. *J. Mol. Graph. Model.* **2011,** *29* (6), 888-893.

53. Oliphant, T. E., Python for Scientific Computing. *Comput. Sci. Eng.* **2007,** *9* (3), 10-20.

54. Millman, K. J.; Jarrod Millman, K.; Aivazis, M., Python for Scientists and Engineers. *Comput. Sci. Eng.* **2011,** *13* (2), 9-12.

55. Kelley, L. A.; Gardner, S. P.; Sutcliffe, M. J., An automated approach for clustering an ensemble of NMR-derived protein structures into conformationally related subfamilies. *Protein Eng.* **1996,** *9* (11), 1063-1065.

56. Humphrey, W.; Dalke, A.; Schulten, K., VMD: Visual molecular dynamics. *J. Mol. Graph.* **1996,** *14* (1), 33-38.

57. Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E., The Protein Data Bank. *Nucleic Acids Res.* **2000,** *28* (1), 235-242.

58. Obermann, W. M.; Sondermann, H.; Russo, A. A.; Pavletich, N. P.; Hartl, F. U., In vivo function of Hsp90 is dependent on ATP binding and ATP hydrolysis. *J. Cell Biol.* **1998,** *143* (4), 901-910.

59. Wright, L.; Barril, X.; Dymock, B.; Sheridan, L.; Surgenor, A.; Beswick, M.; Drysdale, M.; Collier, A.; Massey, A.; Davies, N.; Fink, A.; Fromont, C.; Aherne, W.; Boxall, K.; Sharp, S.; Workman, P.; Hubbard, R. E., Structure-activity relationships in purine-based inhibitor binding to HSP90 isoforms. *Chem. Biol.* **2004,** *11* (6), 775-785.

60. Brough, P. A.; Aherne, W.; Barril, X.; Borgognoni, J.; Boxall, K.; Cansfield, J. E.; Cheung, K.-M. J.; Collins, I.; Davies, N. G. M.; Drysdale, M. J.; Dymock, B.; Eccles, S. A.; Finch, H.; Fink, A.; Hayes, A.; Howes, R.; Hubbard, R. E.; James, K.; Jordan, A. M.; Lockie, A.; Martins, V.; Massey, A.; Matthews, T. P.; McDonald, E.; Northfield, C. J.; Pearl, L. H.; Prodromou, C.; Ray, S.; Raynaud, F. I.; Roughley, S. D.; Sharp, S. Y.; Surgenor, A.; Walmsley, D. L.; Webb, P.; Wood, M.; Workman, P.; Wright, L., 4,5-diarylisoxazole Hsp90 chaperone inhibitors: potential therapeutic agents for the treatment of cancer. *J. Med. Chem.* **2008,** *51* (2), 196-218.

61. Brough, P. A.; Barril, X.; Borgognoni, J.; Chene, P.; Davies, N. G. M.; Davis, B.; Drysdale, M. J.; Dymock, B.; Eccles, S. A.; Garcia-Echeverria, C.; Fromont, C.; Hayes, A.; Hubbard, R. E.; Jordan, A. M.; Jensen, M. R.; Massey, A.; Merrett, A.; Padfield, A.; Parsons, R.; Radimerski, T.; Raynaud, F. I.; Robertson, A.; Roughley, S. D.; Schoepfer, J.; Simmonite, H.; Sharp, S. Y.; Surgenor, A.; Valenti, M.; Walls, S.; Webb, P.; Wood, M.; Workman, P.; Wright, L., Combining hit identification strategies: fragment-based and in silico approaches to orally active 2-aminothieno[2,3-d]pyrimidine inhibitors of the Hsp90 molecular chaperone. *J. Med. Chem.* **2009,** *52* (15), 4794-4809.

62. Murray, C. W.; Carr, M. G.; Callaghan, O.; Chessari, G.; Congreve, M.; Cowan, S.; Coyle, J. E.; Downham, R.; Figueroa, E.; Frederickson, M.; Graham, B.; McMenamin, R.; O'Brien, M. A.; Patel, S.; Phillips, T. R.; Williams, G.; Woodhead, A. J.; Woolford, A. J. A., Fragment-based drug discovery applied to Hsp90. Discovery of two lead series with high ligand efficiency. *J. Med. Chem.* **2010,** *53* (16), 5942-5955.

63. Roughley, S. D.; Hubbard, R. E., How well can fragments explore accessed chemical space? A case study from heat shock protein 90. *J. Med. Chem.* **2011,** *54* (12), 3989-4005.

64. Miura, T.; Fukami, T. A.; Hasegawa, K.; Ono, N.; Suda, A.; Shindo, H.; Yoon, D.-O.; Kim, S.-J.; Na, Y.-J.; Aoki, Y.; Shimma, N.; Tsukuda, T.; Shiratori, Y., Lead generation of heat shock protein 90 inhibitors by a combination of fragment-based approach, virtual screening, and structure-based drug design. *Bioorg. Med. Chem. Lett.* **2011,** *21* (19), 5778-5783.

65. Barta, T. E.; Veal, J. M.; Rice, J. W.; Partridge, J. M.; Fadden, R. P.; Ma, W.; Jenks, M.; Geng, L.; Hanson, G. J.; Huang, K. H.; Barabasz, A. F.; Foley, B. E.; Otto, J.; Hall, S. E., Discovery of benzamide tetrahydro-4H-carbazol-4-ones as novel small molecule inhibitors of Hsp90. *Bioorg. Med. Chem. Lett.* **2008,** *18* (12), 3517-3521.

66. Cho-Schultz, S.; Patten, M. J.; Huang, B.; Elleraas, J.; Gajiwala, K. S.; Hickey, M. J.; Wang, J.; Mehta, P. P.; Kang, P.; Gehring, M. R.; Kung, P.-P.; Sutton, S. C., Solution-phase parallel synthesis of Hsp90 inhibitors. *J. Comb. Chem.* **2009,** *11* (5), 860-874.

67. Kung, P.-P.; Huang, B.; Zhang, G.; Zhou, J. Z.; Wang, J.; Digits, J. A.; Skaptason, J.; Yamazaki, S.; Neul, D.; Zientek, M.; Elleraas, J.; Mehta, P.; Yin, M.-J.; Hickey, M. J.; Gajiwala, K. S.; Rodgers, C.; Davies, J. F.; Gehring, M. R., Dihydroxyphenylisoindoline amides as orally bioavailable inhibitors of the heat shock protein 90 (hsp90) molecular chaperone. *J. Med. Chem.* **2010,** *53* (1), 499-503.

68. Zapf, C. W.; Bloom, J. D.; Li, Z.; Dushin, R. G.; Nittoli, T.; Otteng, M.; Nikitenko, A.; Golas, J. M.; Liu, H.; Lucas, J.; Boschelli, F.; Vogan, E.; Olland, A.; Johnson, M.; Levin, J. I., Discovery of a stable macrocyclic o-aminobenzamide Hsp90 inhibitor which significantly decreases tumor volume in a mouse xenograft model. *Bioorg. Med. Chem. Lett.* **2011,** *21* (15), 4602-4607.

69. Kung, P.-P.; Sinnema, P.-J.; Richardson, P.; Hickey, M. J.; Gajiwala, K. S.; Wang, F.; Huang, B.; McClellan, G.; Wang, J.; Maegley, K.; Bergqvist, S.; Mehta, P. P.; Kania, R., Design strategies to target crystallographic waters applied to the Hsp90 molecular chaperone. *Bioorg. Med. Chem. Lett.* **2011,** *21* (12), 3557-3562.

70. Casale, E.; Amboldi, N.; Brasca, M. G.; Caronni, D.; Colombo, N.; Dalvit, C.; Felder, E. R.; Fogliatto, G.; Galvani, A.; Isacchi, A.; Polucci, P.; Riceputi, L.; Sola, F.; Visco, C.; Zuccotto, F.; Casuscelli, F., Fragment-based hit discovery and structure-based optimization of aminotriazoloquinazolines as novel Hsp90 inhibitors. *Bioorg. Med. Chem.* **2014,** *22* (15), 4135-4150.

71. Davies, N. G. M.; Browne, H.; Davis, B.; Drysdale, M. J.; Foloppe, N.; Geoffrey, S.; Gibbons, B.; Hart, T.; Hubbard, R.; Jensen, M. R.; Mansell, H.; Massey, A.; Matassova, N.; Moore, J. D.; Murray, J.; Pratt, R.; Ray, S.; Robertson, A.; Roughley, S. D.; Schoepfer, J.; Scriven, K.; Simmonite, H.; Stokes, S.; Surgenor, A.; Webb, P.; Wood, M.; Wright, L.; Brough, P., Targeting conserved water molecules: design of 4-aryl-5-cyanopyrrolo[2,3-d]pyrimidine Hsp90 inhibitors using fragment-based screening and structure-based optimization. *Bioorg. Med. Chem.* **2012,** *20* (22), 6770-6789.

72. Chen, D.; Shen, A.; Li, J.; Shi, F.; Chen, W.; Ren, J.; Liu, H.; Xu, Y.; Wang, X.; Yang, X.; Sun, Y.; Yang, M.; He, J.; Wang, Y.; Zhang, L.; Huang, M.; Geng, M.; Xiong, B.; Shen, J., Discovery of potent N-(isoxazol-5-yl)amides as HSP90 inhibitors. *Eur. J. Med. Chem.* **2014,** *87*, 765-781.

73. Ren, J.; Yang, M.; Liu, H.; Cao, D.; Chen, D.; Li, J.; Tang, L.; He, J.; Chen, Y.-L.; Geng, M.; Xiong, B.; Shen, J., Multi-substituted 8-aminoimidazo[1,2-a]pyrazines by Groebke-Blackburn-Bienaymé reaction and their Hsp90 inhibitory activity. *Org. Biomol. Chem.* **2015,** *13* (5), 1531-1535.

74. McBride, C. M.; Levine, B.; Xia, Y.; Bellamacina, C.; Machajewski, T.; Gao, Z.; Renhowe, P.; Antonios-McCrea, W.; Barsanti, P.; Brinner, K.; Costales, A.; Doughan, B.; Lin, X.; Louie, A.; McKenna, M.; Mendenhall, K.; Poon, D.; Rico, A.; Wang, M.; Williams, T. E.; Abrams, T.; Fong, S.; Hendrickson, T.; Lei, D.; Lin, J.; Menezes, D.; Pryer, N.; Taverna, P.; Xu, Y.; Zhou, Y.; Shafer, C. M., Design, structure-activity relationship, and in vivo characterization of the development candidate NVP-HSP990. *J. Med. Chem.* **2014,** *57* (21), 9124-9129.

75. *Gaussian 09*, Gaussian, Inc. Wallingford CT.: 2009.

76. Vanquelef, E.; Simon, S.; Marquant, G.; Garcia, E.; Klimerak, G.; Delepine, J. C.; Cieplak, P.; Dupradeau, F. Y., R.E.D. Server: a web service for deriving RESP and ESP charges and building force field libraries for new molecules and molecular fragments. *Nucleic Acids Res.* **2011,** *39* (suppl), W511-W517.

77. Dupradeau, F.-Y.; Pigache, A.; Zaffran, T.; Savineau, C.; Lelong, R.; Grivel, N.; Lelong, D.; Rosanski, W.; Cieplak, P., The R.E.D. tools: advances in RESP and ESP charge derivation and force field library building. *Phys. Chem. Chem. Phys.* **2010,** *12* (28), 7821.

78. Bayly, C. I.; Cieplak, P.; Cornell, W.; Kollman, P. A., A well-behaved electrostatic potential based method using charge restraints for deriving atomic charges: the RESP model. *J. Phys. Chem.* **1993,** *97* (40), 10269-10280.

79. Kendall, M. G., A NEW MEASURE OF RANK CORRELATION. *Biometrika* **1938,** *30* (1-2), 81-93.

80. Vauquelin, G.; Charlton, S. J., Long-lasting target binding and rebinding as mechanisms to prolong in vivo drug action. *Br. J. Pharmacol.* **2010,** *161* (3), 488-508.

81. Reuben S. Harris, J. P. D., APOBECs and Virus Restriction. *Virology* **2015,** *0*, 131.

82. Kohli, R. M.; Abrams, S. R.; Gajula, K. S.; Maul, R. W.; Gearhart, P. J.; Stivers, J. T., A Portable Hot Spot Recognition Loop Transfers Sequence Preferences from APOBEC Family Members to Activation-induced Cytidine Deaminase. *J. Biol. Chem.* **2009,** *284* (34), 22898.

83. Nabel, C. S.; Lee, J. W.; Wang, L. C.; Kohli, R. M., Nucleic acid determinants for selective deamination of DNA over RNA by activation-induced deaminase. *Proc. Natl. Acad. Sci. U. S. A.* **2013,** *110* (35), 14225.

84. Pang, Y.-P., Novel Zinc Protein Molecular Dynamics Simulations: Steps Toward Antiangiogenesis for Cancer Treatment. *J. Mol. Model.* **1999,** *5* (10), 196-202.

85. McGibbon, R. T.; Beauchamp, K. A.; Harrigan, M. P.; Klein, C.; Swails, J. M.; Hernández, C. X.; Schwantes, C. R.; Wang, L.-P.; Lane, T. J.; Pande, V. S., MDTraj: A Modern Open Library for the Analysis of Molecular Dynamics Trajectories. *Biophys. J.* **2015,** *109* (8), 1528-1532.

86. Hunter, J. D., Matplotlib: A 2D Graphics Environment. *Comput. Sci. Eng.* **2007,** *9* (3), 90-95.

87. Baker, E. N.; Hubbard, R. E., Hydrogen bonding in globular proteins. *Prog. Biophys. Mol. Biol.* **1984,** *44* (2), 97-179.

88. Shi, K.; Kurahashi, K.; Aihara, H., Crystal Structure of the Cancer Genomic DNA Mutator APOBEC3B. **2015**.

89. Shi, K.; Banerjee, S.; Kurahashi, K.; Aihara, H., Crystal Structure of Human APOBEC3A complexed with ssDNA. **2016**.

90. Byeon, I.-J. L.; Ahn, J.; Mitra, M.; Byeon, C.-H.; Hercík, K.; Hritz, J.; Charlton, L. M.; Levin, J. G.; Gronenborn, A. M., NMR structure of human restriction factor APOBEC3A reveals substrate binding and enzyme specificity. *Nat. Commun.* **2013,** *4*, 1890.

91. Kitamura, S.; Ode, H.; Nakashima, M.; Imahashi, M.; Naganawa, Y.; Kurosawa, T.; Yokomaku, Y.; Yamane, T.; Watanabe, N.; Suzuki, A.; Sugiura, W.; Iwatani, Y., The APOBEC3C crystal structure and the interface for HIV-1 Vif binding. *Nat. Struct. Mol. Biol.* **2012,** *19* (10), 1005-1010.

92. Nakashima, M.; Ode, H.; Kawamura, T.; Kitamura, S.; Naganawa, Y.; Awazu, H.; Tsuzuki, S.; Matsuoka, K.; Nemoto, M.; Hachiya, A.; Sugiura, W.; Yokomaku, Y.; Watanabe, N.; Iwatani, Y., Structural Insights into HIV-1 Vif-APOBEC3F Interaction. *J. Virol.* **2015,** *90* (2), 1034-1047.

93. Siu, K. K.; Sultana, A.; Azimi, F. C.; Lee, J. E., Structural determinants of HIV-1 Vif susceptibility and DNA binding in APOBEC3F. *Nat. Commun.* **2013,** *4*, 2593.

94. Bohn, M.-F.; Shandilya, S. M. D.; Albin, J. S.; Kouno, T.; Anderson, B. D.; McDougle, R. M.; Carpenter, M. A.; Rathore, A.; Evans, L.; Davis, A. N.; Zhang, J.; Lu, Y.; Somasundaran, M.; Matsuo, H.; Harris, R. S.; Schiffer, C. A., Crystal structure of the DNA cytosine deaminase APOBEC3F: the catalytically active and HIV-1 Vif-binding domain. *Structure* **2013,** *21* (6), 1042-1050.

95. Shaban, N. M.; Shi, K.; Li, M.; Aihara, H.; Harris, R. S., 1.92 Angstrom Zinc-Free APOBEC3F Catalytic Domain Crystal Structure. *J. Mol. Biol.* **2016,** *428* (11), 2307-2316.

96. Kouno, T.; Luengas, E. M.; Shigematsu, M.; Shandilya, S. M. D.; Zhang, J.; Chen, L.; Hara, M.; Schiffer, C. A.; Harris, R. S.; Matsuo, H., Structure of the Vif-binding domain of the antiviral enzyme APOBEC3G. *Nat. Struct. Mol. Biol.* **2015,** *22* (6), 485-491.

97. Xiao, X.; Li, S.-X.; Yang, H.; Chen, X. S., Crystal structures of APOBEC3G N-domain alone and its complex with DNA. *Nat. Commun.* **2016,** *7*, 12193.

98. Chen, K.-M.; Harjes, E.; Gross, P. J.; Fahmy, A.; Lu, Y.; Shindo, K.; Harris, R. S.; Matsuo, H., Structure of the DNA deaminase domain of the HIV-1 restriction factor APOBEC3G. *Nature* **2008,** *452* (7183), 116-119.

99. Holden, L. G.; Prochnow, C.; Chang, Y. P.; Bransteitter, R.; Chelico, L.; Sen, U.; Stevens, R. C.; Goodman, M. F.; Chen, X. S., Crystal structure of the anti-viral APOBEC3G catalytic domain and functional implications. *Nature* **2008,** *456* (7218), 121-124.

100. Furukawa, A.; Nagata, T.; Matsugami, A.; Habu, Y.; Sugiyama, R.; Hayashi, F.; Kobayashi, N.; Yokoyama, S.; Takaku, H.; Katahira, M., Structure, interaction and real-time monitoring of the enzymatic reaction of wild-type APOBEC3G. *EMBO J.* **2009,** *28* (4), 440-451.

101. Harjes, E.; Gross, P. J.; Chen, K.-M.; Lu, Y.; Shindo, K.; Nowarski, R.; Gross, J. D.; Kotler, M.; Harris, R. S.; Matsuo, H., An extended structure of the APOBEC3G catalytic domain suggests a unique holoenzyme model. *J. Mol. Biol.* **2009,** *389* (5), 819-832.

102. Shandilya, S. M. D.; Nalam, M. N. L.; Nalivaika, E. A.; Gross, P. J.; Valesano, J. C.; Shindo, K.; Li, M.; Munson, M.; Royer, W. E.; Harjes, E.; Kono, T.; Matsuo, H.; Harris, R. S.; Somasundaran, M.; Schiffer, C. A., Crystal structure of the APOBEC3G catalytic domain reveals potential oligomerization interfaces. *Structure* **2010,** *18* (1), 28-38.

103. Li, M.; Shandilya, S. M. D.; Carpenter, M. A.; Rathore, A.; Brown, W. L.; Perkins, A. L.; Harki, D. A.; Solberg, J.; Hook, D. J.; Pandey, K. K.; Parniak, M. A.; Johnson, J. R.; Krogan, N. J.; Somasundaran, M.; Ali, A.; Schiffer, C. A.; Harris, R. S., First-in-class small molecule inhibitors of the single-strand DNA cytosine deaminase APOBEC3G. *ACS Chem. Biol.* **2012,** *7* (3), 506-517.

104. Lu, X.; Zhang, T.; Xu, Z.; Liu, S.; Zhao, B.; Lan, W.; Wang, C.; Ding, J.; Cao, C., Crystal structure of DNA cytidine deaminase ABOBEC3G catalytic deamination domain suggests a binding mode of full-length enzyme to single-stranded DNA. *J. Biol. Chem.* **2015,** *290* (7), 4010-4021.

105. Duarte, J. M.; Srebniak, A.; Schärer, M. A.; Capitani, G., Protein interface classification by evolutionary analysis. *BMC Bioinformatics* **2012,** *13* (1), 334.

106. Chen, J.; Sawyer, N.; Regan, L., Protein–protein interactions: General trends in the relationship between binding affinity and interfacial buried surface area. *Protein Sci.* **2013,** *22* (4), 510.

107. Shi, K.; Carpenter, M. A.; Banerjee, S.; Shaban, N. M.; Kurahashi, K.; Salamango, D. J.; McCann, J. L.; Starrett, G. J.; Duffy, J. V.; Demir, Ö.; Amaro, R. E.; Harki, D. A.; Harris, R. S.;

Aihara, H., Structural basis for targeted DNA cytosine deamination and mutagenesis by APOBEC3A and APOBEC3B. *Nat. Struct. Mol. Biol.* **2016,** *24* (2), 131-139.

108. McDougle, R. M.; Hultquist, J. F.; Stabell, A. C.; Sawyer, S. L.; Harris, R. S., D316 is critical for the enzymatic activity and HIV-1 restriction potential of human and rhesus APOBEC3B. *Virology* **2013,** *441* (1), 31-39.

109. Byeon, I.-J. L.; Byeon, C.-H.; Wu, T.; Mitra, M.; Singer, D.; Levin, J. G.; Gronenborn, A. M., Nuclear Magnetic Resonance Structure of the APOBEC3B Catalytic Domain: Structural Basis for Substrate Binding and DNA Deaminase Activity. *Biochemistry* **2016,** *55* (21), 2944-2959.

110. Harjes, S.; Solomon, W. C.; Li, M.; Chen, K.-M.; Harjes, E.; Harris, R. S.; Matsuo, H., Impact of H216 on the DNA binding and catalytic activities of the HIV restriction factor APOBEC3G. *J. Virol.* **2013,** *87* (12), 7008-7014.

111. Christopoulos, A.; May, L. T.; Avlani, V. A.; Sexton, P. M., G-protein-coupled receptor allosterism: the promise and the problem(s). *Biochem. Soc. Trans.* **2004,** *32* (5), 873--877.

112. Groebe, D. R., Screening for positive allosteric modulators of biological targets. *Drug Discov. Today* **2006,** *11* (13-14), 632--639.

113. May, L. T.; Leach, K.; Sexton, P. M.; Christopoulos, A., Allosteric Modulation of G Protein–Coupled Receptors. *Annu. Rev. Pharmacol. Toxicol.* **2007,** *47* (1), 1--51.

114. Kenakin, T. P., Ligand Detection in the Allosteric World. *J. Biomol. Screen.* **2010,** *15* (2), 119--130.

115. Nussinov, R.; Tsai, C.-J., The Different Ways through Which Specificity Works in Orthosteric and Allosteric Drugs. *Curr. Pharm. Des.* **2012,** *18* (9), 1311--1316.

116. Szilagyi, A.; Nussinov, R.; Csermely, P., Allo-Network Drugs: Extension of the Allosteric Drug Concept to Protein- Protein Interaction and Signaling Networks. *Curr. Top. Med. Chem.* **2013,** *13* (1), 64--77.

117. Grover, A. K., Use of Allosteric Targets in the Discovery of Safer Drugs. *Med. Princ. Pract.* **2013,** *22* (5), 418--426.

118. Ma, B.; Nussinov, R., Druggable Orthosteric and Allosteric Hot Spots to Target Protein-protein Interactions. *Curr. Pharm. Des.* **2014,** *20* (8), 1293--1301.

119. Gunasekaran, K.; Ma, B.; Nussinov, R., Is allostery an intrinsic property of all dynamic proteins? *Proteins* **2004,** *57* (3), 433--43.

120. Wenthur, C. J.; Gentry, P. R.; Mathews, T. P.; Lindsley, C. W., Drugs for Allosteric Sites on Receptors. *Annu. Rev. Pharmacol. Toxicol.* **2014,** *54* (1), 165--184.

121. Groebe, D. R., In search of negative allosteric modulators of biological targets. *Drug Discov. Today* **2009,** *14* (1-2), 41--49.

122. Csermely, P.; Nussinov, R.; Szilagyi, A., Editorial (Hot Topic: From Allosteric Drugs to Allo-Network Drugs: State of the Art and Trends of Design, Synthesis and Computational Methods). *Curr. Top. Med. Chem.* **2013,** *13* (1), 2--4.

123. Muchmore, S. W.; Hajduk, P. J., Crystallography, NMR and virtual screening: integrated tools for drug discovery. *Curr. Opin. Drug Discov. Devel.* **2003,** *6* (4), 544--9.

124. Doman, T. N.; McGovern, S. L.; Witherbee, B. J.; Kasten, T. P.; Kurumbail, R.; Stallings, W. C.; Connolly, D. T.; Shoichet, B. K., Molecular Docking and High-Throughput Screening for Novel Inhibitors of Protein Tyrosine Phosphatase-1B. *J. Med. Chem.* **2002,** *45* (11), 2213--2221.

125. Wood, M. R.; Hopkins, C. R.; Brogan, J. T.; Conn, P. J.; Lindsley, C. W., “Molecular Switches” on mGluR Allosteric Ligands That Modulate Modes of Pharmacology. *Biochemistry* **2011,** *50* (13), 2403--2410.

126. Saalau-bethell, S. M.; Woodhead, A. J.; Chessari, G.; Carr, M. G.; Coyle, J.; Graham, B.; Hiscock, S. D.; Murray, C. W.; Pathuri, P.; Rich, S. J.; Richardson, C. J.; Williams, P. A.; Jhoti, H., Discovery of an allosteric mechanism for the regulation of HCV NS3 protein function. *Nat. Chem. Biol.* **2012,** *8* (11), 920--925.

127. Rawal, R. K.; Murugesan, V.; Katti, S. B., Structure-Activity Relationship Studies on Clinically Relevant HIV-1 NNRTIs. *Curr. Med. Chem.* **2012,** *19* (31), 5364--5380.

128. Reynolds, C.; de Koning, C. B.; Pelly, S. C.; van Otterlo, W. a. L.; Bode, M. L., In search of a treatment for HIV – current therapies and the role of non-nucleoside reverse transcriptase inhibitors (NNRTIs). *Chem. Soc. Rev.* **2012,** *41* (13), 4657.

129. Gentry, P. R.; Sexton, P. M.; Christopoulos, A., Novel Allosteric Modulators of G Protein-coupled Receptors. *J. Biol. Chem.* **2015,** *290* (32), 19478--19488.

130. Regan, M. C.; Romero-Hernandez, A.; Furukawa, H., A structural biology perspective on NMDA receptor pharmacology and function. *Curr. Opin. Struct. Biol.* **2015,** *33*, 68--75.

131. Rutkowski, P.; Lugowska, I.; Kosela-Paterczyk, H.; Kozak, K., Trametinib: a MEK inhibitor for management of metastatic melanoma. *Onco. Targets. Ther.* **2015,** *8*, 2251.

132. Meng, H.; McClendon, C. L.; Dai, Z.; Li, K.; Zhang, X.; He, S.; Shang, E.; Liu, Y.; Lai, L., Discovery of Novel 15-Lipoxygenase Activators To Shift the Human Arachidonic Acid Metabolic Network toward Inflammation Resolution. *J. Med. Chem.* **2015**.

133. Wu, P.; Nielsen, T. E.; Clausen, M. H., FDA-approved small-molecule kinase inhibitors. *Trends Pharmacol. Sci.* **2015,** *36* (7), 422--439.

134. Jazayeri, A.; Dias, J. M.; Marshall, F. H., From G Protein-coupled Receptor Structure Resolution to Rational Drug Design. *J. Biol. Chem.* **2015,** *290* (32), 19489--19495.

135. Tautermann, C. S., GPCR structures in drug design, emerging opportunities with new structures. *Bioorg. Med. Chem. Lett.* **2014,** *24* (17), 4073--4079.

136. Karakas, E.; Furukawa, H., Crystal structure of a heterotetrameric NMDA receptor ion channel. *Science (80-.).* **2014,** *344* (6187), 992--997.

137. Sciara, G.; Mancia, F., Highlights from recently determined structures of membrane proteins: a focus on channels and transporters. *Curr. Opin. Struct. Biol.* **2012,** *22* (4), 476--481.

138. Monod, J.; Wyman, J.; Changeux, J.-P., On the nature of allosteric transitions: A plausible model. *J. Mol. Biol.* **1965,** *12* (1), 88--118.

139. Weber, G., Ligand binding and internal equilibiums in proteins. *Biochemistry* **1972,** *11* (5), 864--878.

140. Cooper, A.; Dryden, D. T., Allostery without conformational change. A plausible model. *Eur. Biophys. J.* **1984,** *11* (2), 103--9.

141. Jardetzky, O., Protein dynamics and conformational transitions in allosteric proteins. *Prog. Biophys. Mol. Biol.* **1996,** *65* (3), 171--219.

142. Kumar, S.; Ma, B.; Tsai, C.-J.; Sinha, N.; Nussinov, R., Folding and binding cascades: Dynamic landscapes and population shifts. *Protein Sci.* **2008,** *9* (1), 10--19.

143. Kern, D.; Zuiderweg, E. R. P., The role of dynamics in allosteric regulation. *Curr. Opin. Struct. Biol.* **2003,** *13* (6), 748--757.

144. Tsai, C.-J.; del Sol, A.; Nussinov, R., Allostery: Absence of a Change in Shape Does Not Imply that Allostery Is Not at Play. *J. Mol. Biol.* **2008,** *378* (1), 1--11.

145. Boehr, D. D.; Nussinov, R.; Wright, P. E., The role of dynamic conformational ensembles in biomolecular recognition. *Nat. Chem. Biol.* **2009,** *5* (11), 789--96.

146. Kar, G.; Keskin, O.; Gursoy, A.; Nussinov, R., Allostery and population shift in drug discovery. *Curr. Opin. Pharmacol.* **2010,** *10* (6), 715--722.

147. Zhou, H.-X., From Induced Fit to Conformational Selection: A Continuum of Binding Mechanism Controlled by the Timescale of Conformational Transitions. *Biophys. J.* **2010,** *98* (6), L15--L17.

148. Nussinov, R.; Ma, B.; Tsai, C. J., Multiple conformational selection and induced fit events take place in allosteric propagation. *Biophys. Chem.* **2014,** *186*, 22--30.

149. Changeux, J.-P., Allostery and the Monod-Wyman-Changeux Model After 50 Years. *Annu. Rev. Biophys.* **2012,** *41* (1), 103--133.

150. Greives, N.; Zhou, H.-X., Both protein dynamics and ligand concentration can shift the binding mechanism between conformational selection and induced fit. *Proc. Natl. Acad. Sci.* **2014,** *111* (28), 10197--202.

151. Guo, J.; Pang, X.; Zhou, H.-X., Two Pathways Mediate Interdomain Allosteric Regulation in Pin1. *Structure* **2015,** *23* (1), 237--247.

152. Motlagh, H. N.; Wrabl, J. O.; Li, J.; Hilser, V. J., The ensemble nature of allostery. *Nature* **2014,** *508* (7496), 331--339.

153. Tsai, C.-J.; del Sol, A.; Nussinov, R., Protein allostery, signal transmission and dynamics: a classification scheme of allosteric mechanisms. *Mol. Biosyst.* **2009,** *5* (3), 207.

154. Kornev, A. P.; Taylor, S. S., Dynamics-Driven Allostery in Protein Kinases. *Trends Biochem. Sci.* **2015,** *40* (11), 628--647.

155. Ma, B.; Tsai, C.-J.; Haliloğlu, T.; Nussinov, R., Dynamic Allostery: Linkers Are Not Merely Flexible. *Structure* **2011,** *19* (7), 907--917.

156. del Sol, A.; Fujihashi, H.; Amoros, D.; Nussinov, R., Residues crucial for maintaining short paths in network communication mediate signaling in proteins. *Mol. Syst. Biol.* **2006,** *2* (1), 2006.0019.

157. del Sol, A.; Tsai, C.-J.; Ma, B.; Nussinov, R., The Origin of Allosteric Functional Modulation: Multiple Pre-existing Pathways. *Structure* **2009,** *17* (8), 1042--1050.

158. Malmstrom, R. D.; Kornev, A. P.; Taylor, S. S.; Amaro, R. E., Allostery through the computational microscope: cAMP activation of a canonical signalling domain. *Nat. Commun.* **2015,** *6*, 7588.

159. Pontiggia, F.; Pachov, D. V.; Clarkson, M. W.; Villali, J.; Hagan, M. F.; Pande, V. S.; Kern, D., Free energy landscape of activation in a signalling protein at atomic resolution. *Nat. Commun.* **2015,** *6*, 7284.

160. Hebeisen, P.; Haap, W.; Kuhn, B.; Mohr, P.; Wessel, H. P.; Zutter, U.; Kirchner, S.; Ruf, A.; Benz, J.; Joseph, C.; Alvarez-Snchez, R.; Gubler, M.; Schott, B.; Benardeau, A.; Tozzo, E.; Kitas, E., Orally active aminopyridines as inhibitors of tetrameric fructose-1,6-bisphosphatase. *Bioorg. Med. Chem. Lett.* **2011,** *21* (11), 3237--3242.

161. Zarzycki, M.; Kołodziejczyk, R.; Maciaszczyk-Dziubinska, E.; Wysocki, R.; Jaskolski, M.; Dzugaj, A., Structure of E69Q mutant of human muscle fructose-1,6-bisphosphatase. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **2011,** *67* (12), 1028--1034.

162. Nussinov, R.; Tsai, C.-J.; Csermely, P., Allo-network drugs: harnessing allostery in cellular networks. *Trends Pharmacol. Sci.* **2011,** *32* (12), 686--693.

163. Gilmartin, A. G.; Bleam, M. R.; Groy, A.; Moss, K. G.; Minthorn, E. A.; Kulkarni, S. G.; Rominger, C. M.; Erskine, S.; Fisher, K. E.; Yang, J.; Zappacosta, F.; Annan, R.; Sutton, D.;

Laquerre, S. G., GSK1120212 (JTP-74057) is an inhibitor of MEK activity and activation with favorable pharmacokinetic properties for sustained in vivo pathway inhibition. *Clin. Cancer Res.* **2011,** *17* (5), 989--1000.

164. Sebolt-Leopold, J. S.; Dudley, D. T.; Herrera, R., Blockade of the MAP kinase pathway suppresses growth of colon tumors in vivo. *Nat. Med.* **1999,** *5* (7), 810--816.

165. Ohren, J. F.; Chen, H.; Pavlovsky, A.; Whitehead, C.; Zhang, E.; Kuffa, P.; Yan, C.; McConnell, P.; Spessard, C.; Banotai, C.; Mueller, W. T.; Delaney, A.; Omer, C.; Sebolt-Leopold, J.; Dudley, D. T.; Leung, I. K.; Flamme, C.; Warmus, J.; Kaufman, M.; Barrett, S.; Tecle, H.; Hasemann, C. a., Structures of human MAP kinase kinase 1 (MEK1) and MEK2 describe novel noncompetitive kinase inhibition. *Nat. Struct. Mol. Biol.* **2004,** *11* (12), 1192--1197.

166. Zhao, Y.; Adjei, A. A., The clinical development of MEK inhibitors. *Nat. Publ. Gr.* **2014,** *11* (7), 385--400.

167. Hackos, D. H.; Lupardus, P. J.; Grand, T.; Sheng, M.; Zhou, Q.; Hanson, J. E.; Chen, Y.; Wang, T.-m.; Reynen, P., Positive Allosteric Modulators of GluN2A-Containing NMDARs with Distinct Modes of Action and Impacts on Circuit Function Article Positive Allosteric Modulators of GluN2A-Containing NMDARs with Distinct Modes of Action and Impacts on Circuit Function. *Neuron* **2016,** *89* (5), 1--17.

168. Talavera, D.; Lovell, S. C.; Whelan, S., Covariation Is a Poor Measure of Molecular Coevolution. *Mol. Biol. Evol.* **2015,** *32* (9), 2456--2468.

169. Fodor, A. A.; Aldrich, R. W., On Evolutionary Conservation of Thermodynamic Coupling in Proteins. *J. Biol. Chem.* **2004,** *279* (18), 19046--19050.

170. Fodor, A. A.; Aldrich, R. W., Influence of conservation on calculations of amino acid covariance in multiple sequence alignments. *Proteins Struct. Funct. Bioinforma.* **2004,** *56* (2), 211--221.

171. Johansson, F.; Toh, H., Relative von Neumann entropy for evaluating amino acid conservation. *J. Bioinform. Comput. Biol.* **2010,** *8* (5), 809--23.

172. de Juan, D.; Pazos, F.; Valencia, A., Emerging methods in protein co-evolution. *Nat. Rev. Genet.* **2013,** *14* (4), 249--261.

173. Shannon, C. E., A Mathematical Theory of Communication. *Bell Syst. Tech. J.* **1948,** *27* (3), 379--423.

174. Shenkin, P. S.; Erman, B.; Mastrandrea, L. D., Information-theoretical entropy as a measure of sequence variability. *Proteins Struct. Funct. Genet.* **1991,** *11* (4), 297--313.

175. Vinga, S., Information theory applications for biological sequence analysis. *Brief. Bioinform.* **2014,** *15* (3), 376--389.

176. Kullback, S.; Leibler, R. A., On Information and Sufficiency. *Ann. Math. Stat.* **1951,** *22* (1), 79--86.

177. Zhang, S. W.; Zhang, Y. L.; Pan, Q.; Cheng, Y. M.; Chou, K. C., Estimating residue evolutionary conservation by introducing von Neumann entropy and a novel gap-treating approach. *Amino Acids* **2008,** *35* (2), 495--501.

178. Lichtarge, O.; Bourne, H. R.; Cohen, F. E., An evolutionary trace method defines binding surfaces common to protein families. *J. Mol. Biol.* **1996,** *257* (2), 342--58.

179. Madabushi, S.; Gross, A. K.; Philippi, A.; Meng, E. C.; Wensel, T. G.; Lichtarge, O., Evolutionary Trace of G Protein-coupled Receptors Reveals Clusters of Residues That Determine Global and Class-specific Functions. *J. Biol. Chem.* **2004,** *279* (9), 8126--8132.

180. Raviscioni, M.; He, Q.; Salicru, E. M.; Smith, C. L.; Lichtarge, O., Evolutionary identification of a subtype specific functional site in the ligand binding domain of steroid receptors. *Proteins Struct. Funct. Bioinforma.* **2006,** *64* (4), 1046--1057.

181. Rodriguez, G. J.; Yao, R.; Lichtarge, O.; Wensel, T. G., Evolution-guided discovery and recoding of allosteric pathway specificity determinants in psychoactive bioamine receptors. *Proc. Natl. Acad. Sci.* **2010,** *107* (17), 7787--7792.

182. Mihalek, I.; Re, A Family of Evolution–Entropy Hybrid Methods for Ranking Protein Residues by Importance. *J. Mol. Biol.* **2004,** *336* (5), 1265--1282.

183. Mihalek, I.; Res, I.; Lichtarge, O., Evolutionary trace report _ maker: a new type of service for comparative analysis of proteins. *Bioinformatics* **2006,** *22* (13), 1656--1657.

184. Merkl, R.; Zwick, M., H2r: Identification of evolutionary important residues by means of an entropy based analysis of multiple sequence alignments. *BMC Bioinformatics* **2008,** *9* (1), 151.

185. Janda, J.-O.; Popal, A.; Bauer, J.; Busch, M.; Klocke, M.; Spitzer, W.; Keller, J.; Merkl, R., H2rs: Deducing evolutionary and functionally important residue positions by means of an entropy and similarity based analysis of multiple sequence alignments. *BMC Bioinformatics* **2014,** *15* (1), 118.

186. Pel, Comparative analysis of sequence covariation methods to mine evolutionary hubs: Examples from selected GPCR families. *Proteins Struct. Funct. Bioinforma.* **2014,** *82* (9), 2141--2156.

187. Brown, C. A.; Brown, K. S., Validation of Coevolving Residue Algorithms via Pipeline Sensitivity Analysis: ELSC and OMES and ZNMI, Oh My! *PLoS One* **2010,** *5* (6), e10779.

188. Simonetti, F. L.; Teppa, E.; Chernomoretz, A.; Nielsen, M., MISTIC: mutual information server to infer coevolution. *Nucleic Acids Res.* **2013,** *41* (W1), W8--W14.

189. Horner, D. S.; Pirovano, W.; Pesole, G., Correlated substitution analysis and the prediction of amino acid structural contacts. *Brief. Bioinform.* **2007,** *9* (1), 46--56.

190. Gao, H.; Dou, Y.; Yang, J.; Wang, J., New methods to measure residues coevolution in proteins. *BMC Bioinformatics* **2011,** *12* (1), 206.

191. Dunn, S. D.; Wahl, L. M.; Gloor, G. B., Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction. *Bioinformatics* **2008,** *24* (3), 333--340.

192. Wollenberg, K. R.; Atchley, W. R., Separation of phylogenetic and functional associations in biological sequences by using the parametric bootstrap. *Proc. Natl. Acad. Sci.* **2000,** *97* (7), 3288--3291.

193. Buslje, C. M.; Santos, J.; Delfino, J. M.; Nielsen, M., Correction for phylogeny, small number of observations and data redundancy improves the identification of coevolving amino acid pairs using mutual information. *Bioinformatics* **2009,** *25* (9), 1125--1131.

194. Clark, G. W.; Ackerman, S. H.; Tillier, E. R.; Gatti, D. L., Multidimensional mutual information methods for the analysis of covariation in multiple sequence alignments. *BMC Bioinformatics* **2014,** *15* (1), 157.

195. Martin, L. C.; Gloor, G. B.; Dunn, S. D.; Wahl, L. M., Using information theory to search for co-evolving residues in proteins. *Bioinformatics* **2005,** *21* (22), 4116--4124.

196. Crooks, G. E.; Wolfe, J.; Brenner, S. E., Measurements of protein sequence-structure correlations. *Proteins Struct. Funct. Genet.* **2004,** *57* (4), 804--810.

197. Olmea, O.; Valencia, A., Improving contact predictions by the combination of correlated mutations and other sources of sequence information. *Fold. Des.* **1997,** *2*, S25--S32.

198. McLachlan, A. D., Tests for comparing related amino-acid sequences. Cytochrome c and cytochrome c551. *J. Mol. Biol.* **1971,** *61* (2), 409--424.

199. Halperin, I.; Wolfson, H.; Nussinov, R., Correlated mutations: Advances and limitations. A study on fusion proteins and on the Cohesin-Dockerin families. *Proteins Struct. Funct. Bioinforma.* **2006,** *63* (4), 832--845.

200. Kass, I.; Horovitz, A., Mapping pathways of allosteric communication in GroEL by analysis of correlated mutations. *Proteins Struct. Funct. Genet.* **2002,** *48* (4), 611--617.

201. Ackerman, S. H.; Tillier, E. R.; Gatti, D. L., Accurate Simulation and Detection of Coevolution Signals in Multiple Sequence Alignments. *PLoS One* **2012,** *7* (10), e47108.

202. Lockless, S. W., Evolutionarily Conserved Pathways of Energetic Connectivity in Protein Families. *Science (80-.).* **1999,** *286* (5438), 295--299.

203. Ranganathan, R.; Rivoire, O. *Note 109: A summary of SCA calculations*; 2012; pp 1--11.

204. Sel, G. M.; Lockless, S. W.; Wall, M. A.; Ranganathan, R., Evolutionarily conserved networks of residues mediate allosteric communication in proteins. *Nat. Struct. Biol.* **2003,** *10* (1), 59--69.

205. Halabi, N.; Rivoire, O.; Leibler, S.; Ranganathan, R., Protein Sectors: Evolutionary Units of Three-Dimensional Structure. *Cell* **2009,** *138* (4), 774--786.

206. Reynolds, K. A.; McLaughlin, R. N.; Ranganathan, R., Hot Spots for Allosteric Regulation on Protein Surfaces. *Cell* **2011,** *147* (7), 1564--1575.

207. Russ, W. P.; Lowery, D. M.; Mishra, P.; Yaffe, M. B.; Ranganathan, R., Natural-like function in artificial WW domains. *Nature* **2005,** *437* (7058), 579--583.

208. Socolich, M.; Lockless, S. W.; Russ, W. P.; Lee, H.; Gardner, K. H.; Ranganathan, R., Evolutionary information for specifying a protein fold. *Nature* **2005,** *437* (7058), 512--518.

209. Baths, V.; Roy, U., Identification of distant co-evolving residues in antigen 85C from Mycobacterium tuberculosis using statistical coupling analysis of the esterase family proteins. *J. Biomed. Res.* **2011,** *25* (3), 165--169.

210. The spatial architecture of protein function and adaptation. *Nature* **2012,** *491* (7422), 138--142.

211. Peterson, F. C.; Penkert, R. R.; Volkman, B. F.; Prehoda, K. E., Cdc42 Regulates the Par-6 PDZ Domain through an Allosteric CRIB-PDZ Transition. *Mol. Cell* **2004,** *13* (5), 665--676.

212. Hatley, M. E.; Lockless, S. W.; Gibson, S. K.; Gilman, A. G.; Ranganathan, R., Allosteric determinants in guanine nucleotide-binding proteins. *Proc. Natl. Acad. Sci.* **2003,** *100* (24), 14445--14450.

213. Shulman, A. I.; Larson, C.; Mangelsdorf, D. J.; Ranganathan, R., Structural Determinants of Allosteric Ligand Activation in RXR Heterodimers. *Cell* **2004,** *116* (3), 417--429.

214. Smock, R. G.; Rivoire, O.; Russ, W. P.; Swain, J. F.; Leibler, S.; Ranganathan, R.; Gierasch, L. M., An interdomain sector mediating allostery in Hsp70 molecular chaperones. *Mol. Syst. Biol.* **2010,** *6*, 414.

215. Chi, C. N.; Elfstrom, L.; Shi, Y.; Snall, T.; Engstrom, A.; Jemth, P., Reassessing a sparse energetic network within a single protein domain. *Proc. Natl. Acad. Sci.* **2008,** *105* (12), 4679--4684.

216. Fuentes, E. J.; Der, C. J.; Lee, A. L., Ligand-dependent dynamics and intramolecular signaling in a PDZ domain. *J. Mol. Biol.* **2004,** *335* (4), 1105--15.

217. Dima, R. I.; Thirumalai, D., Proteins associated with diseases show enhanced sequence correlation between charged residues. *Bioinformatics* **2004,** *20* (15), 2345--2354.

218. Dima, R. I., Determination of network of residues that regulate allostery in protein families using sequence analysis. *Protein Sci.* **2006,** *15* (2), 258--268.

219. O'Donovan, C., High-quality protein knowledge resource: SWISS-PROT and TrEMBL. *Brief. Bioinform.* **2002,** *3* (3), 275--284.

220. Boeckmann, B., The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.* **2003,** *31* (1), 365--370.

221. Getz, G.; Levine, E.; Domany, E., Coupled two-way clustering analysis of gene microarray data. *Proc. Natl. Acad. Sci.* **2000,** *97* (22), 12079--12084.

222. Dekker, J. P.; Fodor, A.; Aldrich, R. W.; Yellen, G., A perturbation-based method for calculating explicit likelihood of evolutionary co-variance in multiple sequence alignments. *Bioinformatics* **2004,** *20* (10), 1565--1572.

223. Yip, K. Y.; Patel, P.; Kim, P. M.; Engelman, D. M.; McDermott, D.; Gerstein, M., An integrated system for studying residue coevolution in proteins. *Bioinformatics* **2008,** *24* (2), 290--292.

224. Kozma, D.; Simon, I.; Tusnady, G. E., CMWeb: an interactive on-line tool for analysing residue-residue contacts and contact prediction methods. *Nucleic Acids Res.* **2012,** *40* (W1), W329--W333.

225. Weigt, M.; White, R. A.; Szurmant, H.; Hoch, J. A.; Hwa, T., Identification of direct residue contacts in protein-protein interaction by message passing. *Proc. Natl. Acad. Sci.* **2009,** *106* (1), 67--72.

226. Morcos, F.; Pagnani, A.; Lunt, B.; Bertolino, A.; Marks, D. S.; Sander, C.; Zecchina, R.; Onuchic, J. N.; Hwa, T.; Weigt, M., Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proc. Natl. Acad. Sci.* **2011,** *108* (49), E1293--E1301.

227. Schug, A.; Weigt, M.; Onuchic, J. N.; Hwa, T.; Szurmant, H., High-resolution protein complexes from integrating genomic information with molecular simulation. *Proc. Natl. Acad. Sci.* **2009,** *106* (52), 22124--22129.

228. Sulkowska, J. I.; Morcos, F.; Weigt, M.; Hwa, T.; Onuchic, J. N., Genomics-aided structure prediction. *Proc. Natl. Acad. Sci.* **2012,** *109* (26), 10340--10345.

229. Szurmant, H.; Hoch, J. A., Statistical analyses of protein sequence alignments identify structures and mechanisms in signal activation of sensor histidine kinases. *Mol. Microbiol.* **2013,** *87* (4), 707--712.

230. Procaccini, A.; Lunt, B.; Szurmant, H.; Hwa, T.; Weigt, M., Dissecting the Specificity of Protein-Protein Interaction in Bacterial Two-Component Signaling: Orphans and Crosstalks. *PLoS One* **2011,** *6* (5), e19729.

231. Morcos, F.; Jana, B.; Hwa, T.; Onuchic, J. N., Coevolutionary signals across protein lineages help capture multiple protein conformations. *Proc. Natl. Acad. Sci.* **2013,** *110* (51), 20533--20538.

232. Jones, D. T.; Buchan, D. W. A.; Cozzetto, D.; Pontil, M., PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. *Bioinformatics* **2012,** *28* (2), 184--190.

233. Eckmann, J. P.; Kamphorst, S. O.; Ruelle, D., Recurrence Plots of Dynamical Systems. *Europhys. Lett.* **1987,** *4* (9), 973--977.

234. Zbilut, J. P.; Giuliani, A.; Webber, C. L.; Colosimo, A., Recurrence quantification analysis in structure-function relationships of proteins: an overview of a general methodology applied to the case of TEM-1 beta-lactamase. *Protein Eng. Des. Sel.* **1998,** *11* (2), 87--93.

235. Namboodiri, S.; Verma, C.; Dhar, P. K.; Giuliani, A.; Nair, A. S., Sequence signatures of allosteric proteins towards rational design. *Syst. Synth. Biol.* **2010,** *4* (4), 271--280.

236. Porrello, A.; Soddu, S.; Zbilut, J. P.; Crescenzi, M.; Giuliani, A., Discrimination of single amino acid mutations of the p53 protein by means of deterministic singularities of recurrence quantification analysis. *Proteins Struct. Funct. Bioinforma.* **2004,** *55* (3), 743--755.

237. Giuliani, A.; Tomasi, M., Recurrence quantification analysis reveals interaction partners in paramyxoviridae envelope glycoproteins. *Proteins Struct. Funct. Genet.* **2002,** *46* (2), 171--176.

238. Colafranceschi, M.; Colosimo, A.; Zbilut, J. P.; Uversky, V. N.; Giuliani, A., Structure-related statistical singularities along protein sequences: A correlation study. *J. Chem. Inf. Model.* **2005,** *45* (1), 183--189.

239. Namboodiri, S.; Giuliani, A.; Nair, A. S.; Dhar, P. K., Looking for a sequence based allostery definition: a statistical journey at different resolution scales. *J. Theor. Biol.* **2012,** *304*, 211--8.

240. Zbilut, J. P.; Sirabella, P.; Giuliani, A.; Manetti, C.; Colosimo, A.; Webber, C. L., Review of Nonlinear Analysis of Proteins Through Recurrence Quantification. *Cell Biochem. Biophys.* **2002,** *36* (1), 67--88.

241. Rivoire, O., Elements of Coevolution in Biological Sequences. *Phys. Rev. Lett.* **2013,** *110* (17), 178102.

242. Mao, W.; Kaya, C.; Dutta, A.; Horovitz, A.; Bahar, I., Comparative study of the effectiveness and limitations of current methods for detecting sequence coevolution. *Bioinformatics* **2015,** *31* (12), 1929--1937.

243. Bakan, A.; Meireles, L. M.; Bahar, I., ProDy: Protein Dynamics Inferred from Theory and Experiments. *Bioinformatics* **2011,** *27* (11), 1575--1577.

244. Gbel, U.; Sander, C.; Schneider, R.; Valencia, A., Correlated mutations and residue contacts in proteins. *Proteins Struct. Funct. Genet.* **1994,** *18* (4), 309--317.

245. Fares, M. A., A Novel Method for Detecting Intramolecular Coevolution: Adding a Further Dimension to Selective Constraints Analyses. *Genetics* **2006,** *173* (1), 9--23.

246. Fares, M. A.; McNally, D., CAPS: coevolution analysis using protein sequences. *Bioinformatics* **2006,** *22* (22), 2821--2822.

247. Iserte, J.; Simonetti, F. L.; Zea, D. J.; Teppa, E.; Marino-Buslje, C., I-COMS: Interprotein-COrrelated Mutations Server. *Nucleic Acids Res.* **2015,** *43* (W1), W320--W325.

248. Jeong, C. S.; Kim, D., Reliable and robust detection of coevolving protein residues. *Protein Eng. Des. Sel.* **2012,** *25* (11), 705--713.

249. Berezin, C.; Glaser, F.; Rosenberg, J.; Paz, I.; Pupko, T.; Fariselli, P.; Casadio, R.; Ben-Tal, N., ConSeq: the identification of functionally and structurally important residues in protein sequences. *Bioinformatics* **2004,** *20* (8), 1322--1324.

250. Glaser, F.; Pupko, T.; Paz, I.; Bell, R. E.; Bechor-Shental, D.; Martz, E.; Ben-Tal, N., ConSurf: Identification of Functional Regions in Proteins by Surface-Mapping of Phylogenetic Information. *Bioinformatics* **2003,** *19* (1), 163--164.

251. Landau, M.; Mayrose, I.; Rosenberg, Y.; Glaser, F.; Martz, E.; Pupko, T.; Ben-Tal, N., ConSurf 2005: the projection of evolutionary conservation scores of residues on protein structures. *Nucleic Acids Res.* **2005,** *33* (Web Server issue), W299--W302.

252. Armon, A.; Graur, D.; Ben-Tal, N., ConSurf: an algorithmic tool for the identification of functional regions in proteins by surface mapping of phylogenetic information. *J. Mol. Biol.* **2001,** *307* (1), 447--63.

253. Ashkenazy, H.; Erez, E.; Martz, E.; Pupko, T.; Ben-Tal, N., ConSurf 2010: calculating evolutionary conservation in sequence and structure of proteins and nucleic acids. *Nucleic Acids Res.* **2010,** *38* (Web Server), W529--W533.

254. Celniker, G.; Nimrod, G.; Ashkenazy, H.; Glaser, F.; Martz, E.; Mayrose, I.; Pupko, T.; Ben-Tal, N., ConSurf: Using Evolutionary Data to Raise Testable Hypotheses about Protein Function. *Isr. J. Chem.* **2013,** *53* (3-4), 199--206.

255. Liu, Y.; Gierasch, L. M.; Bahar, I., Role of Hsp70 ATPase Domain Intrinsic Dynamics and Sequence Evolution in Enabling its Functional Interactions with NEFs. *PLoS Comput. Biol.* **2010,** *6* (9), e1000931.

256. Liu, Y.; Bahar, I., Sequence Evolution Correlates with Structural Dynamics. *Mol. Biol. Evol.* **2012,** *29* (9), 2253--2263.

257. Skjrven, L.; Yao, X.-Q.; Scarabelli, G.; Grant, B. J., Integrating protein structural dynamics and evolutionary analysis with Bio3D. *BMC Bioinformatics* **2014,** *15* (1), 399.

258. Grant, B. J.; Rodrigues, A. P. C.; ElSawy, K. M.; McCammon, J. A.; Caves, L. S. D., Bio3d: an R package for the comparative analysis of protein structures. *Bioinformatics* **2006,** *22* (21), 2695--2696.

259. Li, Z.; Huang, Y.; Ouyang, Y.; Jiao, Y.; Xing, H.; Liao, L.; Jiang, S.; Shao, Y.; Ma, L., CorMut: an R/Bioconductor package for computing correlated mutations based on selection pressure. *Bioinformatics* **2014,** *30* (14), 2073--2075.

260. Rhee, S.-Y.; Liu, T. F.; Holmes, S. P.; Shafer, R. W., HIV-1 Subtype B Protease and Reverse Transcriptase Amino Acid Covariation. *PLoS Comput. Biol.* **2007,** *3* (5), e87.

261. Hurst, L. D., The Ka/Ks ratio: diagnosing the form of sequence evolution. *Trends Genet.* **2002,** *18* (9), 486--487.

262. Sliwoski, G.; Kothiwale, S.; Meiler, J.; Lowe, E. W., Computational Methods in Drug Discovery. *Pharmacol. Rev.* **2014,** *66* (1), 334-395.

263. Amaro, R. E.; Baron, R.; Andrew McCammon, J., An improved relaxed complex scheme for receptor flexibility in computer-aided drug design. *J. Comput. Aided Mol. Des.* **2008,** *22* (9), 693-705.

264. Doman, T. N.; McGovern, S. L.; Witherbee, B. J.; Kasten, T. P.; Kurumbail, R.; Stallings, W. C.; Connolly, D. T.; Shoichet, B. K., Molecular docking and high-throughput screening for novel inhibitors of protein tyrosine phosphatase-1B. *J. Med. Chem.* **2002,** *45* (11), 2213-2221.

265. Yuriev, E.; Holien, J.; Ramsland, P. A., Improvements, trends, and new ideas in molecular docking: 2012-2013 in review. *J. Mol. Recognit.* **2015,** *28* (10), 581-604.

266. Guedes, I. A.; de Magalhães, C. S.; Dardenne, L. E., Receptor-ligand molecular docking. *Biophys. Rev.* **2014,** *6* (1), 75-87.

267. Cheng, T.; Li, Q.; Zhou, Z.; Wang, Y.; Bryant, S. H., Structure-Based Virtual Screening for Drug Discovery: a Problem-Centric Review. *AAPS J.* **2012,** *14* (1), 133.

268. Abel, R.; Wang, L.; Mobley, D. L.; Friesner, R. A., A Critical Review of Validation, Blind Testing, and Real- World Use of Alchemical Protein-Ligand Binding Free Energy Calculations. *Curr. Top. Med. Chem.* **2017,** *17* (23).

269. Grinter, S. Z.; Zou, X., Challenges, Applications, and Recent Advances of Protein-Ligand Docking in Structure-Based Drug Design. *Molecules* **2014,** *19* (7), 10150-10176.

270. Warren, G. L.; Andrews, C. W.; Capelli, A.-M.; Clarke, B.; LaLonde, J.; Lambert, M. H.; Lindvall, M.; Nevins, N.; Semus, S. F.; Senger, S.; Tedesco, G.; Wall, I. D.; Woolven, J. M.; Peishoff, C. E.; Head, M. S., A critical assessment of docking programs and scoring functions. *J. Med. Chem.* **2006,** *49* (20), 5912-5931.

271. Jain, A. N., Bias, reporting, and sharing: computational evaluations of docking methods. *J. Comput. Aided Mol. Des.* **2008,** *22* (3-4), 201-212.

272. Mobley, D. L.; Gilson, M. K., Predicting Binding Free Energies: Frontiers and Benchmarks. *Annu. Rev. Biophys.* **2017,** *46*, 531-558.

273. Carlson, H. A., Lessons Learned over Four Benchmark Exercises from the Community Structure-Activity Resource. *J. Chem. Inf. Model.* **2016,** *56* (6), 951-954.

274. Carlson, H. A.; Smith, R. D.; Damm-Ganamet, K. L.; Stuckey, J. A.; Ahmed, A.; Convery, M. A.; Somers, D. O.; Kranz, M.; Elkins, P. A.; Cui, G.; Peishoff, C. E.; Lambert, M. H.; Dunbar, J. B., Jr., CSAR 2014: A Benchmark Exercise Using Unpublished Data from Pharma. *J. Chem. Inf. Model.* **2016,** *56* (6), 1063-1077.

275. Smith, R. D.; Damm-Ganamet, K. L.; Dunbar, J. B., Jr.; Ahmed, A.; Chinnaswamy, K.; Delproposto, J. E.; Kubish, G. M.; Tinberg, C. E.; Khare, S. D.; Dou, J.; Doyle, L.; Stuckey, J. A.; Baker, D.; Carlson, H. A., CSAR Benchmark Exercise 2013: Evaluation of Results from a Combined Computational Protein Design, Docking, and Scoring/Ranking Challenge. *J. Chem. Inf. Model.* **2016,** *56* (6), 1022-1031.

276. Damm-Ganamet, K. L.; Smith, R. D.; Dunbar, J. B., Jr.; Stuckey, J. A.; Carlson, H. A., CSAR benchmark exercise 2011-2012: evaluation of results from docking and relative ranking of blinded congeneric series. *J. Chem. Inf. Model.* **2013,** *53* (8), 1853-1870.

277. Smith, R. D.; Dunbar, J. B., Jr.; Ung, P. M.-U.; Esposito, E. X.; Yang, C.-Y.; Wang, S.; Carlson, H. A., CSAR benchmark exercise of 2010: combined evaluation across all submitted scoring functions. *J. Chem. Inf. Model.* **2011,** *51* (9), 2115-2131.

278. Moult, J.; Fidelis, K.; Kryshtafovych, A.; Schwede, T.; Tramontano, A., Critical assessment of methods of protein structure prediction (CASP)--round x. *Proteins* **2014,** *82 Suppl 2*, 1-6.

279. Kryshtafovych, A.; Monastyrskyy, B.; Fidelis, K., CASP prediction center infrastructure and evaluation measures in CASP10 and CASP ROLL. *Proteins* **2014,** *82 Suppl 2*, 7-13.

280. Taylor, T. J.; Tai, C.-H.; Huang, Y. J.; Block, J.; Bai, H.; Kryshtafovych, A.; Montelione, G. T.; Lee, B., Definition and classification of evaluation units for CASP10. *Proteins* **2014,** *82 Suppl 2*, 14-25.

281. Kufareva, I.; Rueda, M.; Katritch, V.; Stevens, R. C.; Abagyan, R., Status of GPCR Modeling and Docking as Reflected by Community-wide GPCR Dock 2010 Assessment. *Structure* **2011,** *19* (8), 1108-1126.

282. Kufareva, I.; Katritch, V.; Stevens, R. C.; Abagyan, R., Advances in GPCR Modeling Evaluated by the GPCR Dock 2013 Assessment: Meeting New Challenges. *Structure* **2014,** *22* (8), 1120-1139.

283. Michino, M.; Abola, E.; Brooks, C. L.; Scott Dixon, J.; Moult, J.; Stevens, R. C., Community-wide assessment of GPCR structure modelling and ligand docking: GPCR Dock 2008. *Nat. Rev. Drug Discov.* **2009,** *8* (6), 455-463.

284. Haas, J.; Barbato, A.; Behringer, D.; Studer, G.; Roth, S.; Bertoni, M.; Mostaguir, K.; Gumienny, R.; Schwede, T., Continuous Automated Model Evaluation (CAMEO) Complementing the Critical Assessment of Structure Prediction in CASP12. *Proteins* **2017**.

285. Bannan, C. C.; Burley, K. H.; Chiu, M.; Shirts, M. R.; Gilson, M. K.; Mobley, D. L., Blind prediction of cyclohexane-water distribution coefficients from the SAMPL5 challenge. *J. Comput. Aided Mol. Des.* **2016,** *30* (11), 927-944.

286. Gathiaka, S.; Liu, S.; Chiu, M.; Yang, H.; Stuckey, J. A.; Kang, Y. N.; Delproposto, J.; Kubish, G.; Dunbar, J. B., Jr.; Carlson, H. A.; Burley, S. K.; Walters, W. P.; Amaro, R. E.; Feher, V. A.; Gilson, M. K., D3R grand challenge 2015: Evaluation of protein-ligand pose and affinity predictions. *J. Comput. Aided Mol. Des.* **2016,** *30* (9), 651-668.

287. Gaieb, Z.; Liu, S.; Gathiaka, S.; Chiu, M.; Yang, H.; Shao, C.; Feher, V. A.; Patrick Walters, W.; Kuhn, B.; Rudolph, M. G.; Burley, S. K.; Gilson, M. K.; Amaro, R. E., D3R Grand Challenge 2: blind prediction of protein–ligand poses, affinity rankings, and relative binding free energies. *J. Comput. Aided Mol. Des.* **2017**, 1-20.

288. Shamsara, J., CrossDocker: a tool for performing cross-docking using Autodock Vina. *Springerplus* **2016,** *5*.

289. Kumar, A.; Zhang, K. Y. J., A cross docking pipeline for improving pose prediction and virtual screening performance. *J. Comput. Aided Mol. Des.* **2017**.

290. Sastry, G. M.; Adzhigirey, M.; Day, T.; Annabhimoju, R.; Sherman, W., Protein and ligand preparation: parameters, protocols, and influence on virtual screening enrichments. *J. Comput. Aided Mol. Des.* **2013,** *27* (3), 221-234.

291. Dolinsky, T. J.; Czodrowski, P.; Li, H.; Nielsen, J. E.; Jensen, J. H.; Klebe, G.; Baker, N. A., PDB2PQR: expanding and upgrading automated preparation of biomolecular structures for molecular simulations. *Nucleic Acids Res.* **2007,** *35* (Web Server issue), W522-5.

292. Irwin, J. J.; Shoichet, B. K., ZINC − A Free Database of Commercially Available Compounds for Virtual Screening. *J. Chem. Inf. Model.* **2005,** *45* (1), 177-182.

293. Rose, P. W.; Prlić, A.; Altunkaya, A.; Bi, C.; Bradley, A. R.; Christie, C. H.; Costanzo, L. D.; Duarte, J. M.; Dutta, S.; Feng, Z.; Green, R. K.; Goodsell, D. S.; Hudson, B.; Kalro, T.; Lowe, R.; Peisach, E.; Randle, C.; Rose, A. S.; Shao, C.; Tao, Y.-P.; Valasatava, Y.; Voigt, M.; Westbrook, J. D.; Woo, J.; Yang, H.; Young, J. Y.; Zardecki, C.; Berman, H. M.; Burley, S. K., The RCSB protein data bank: integrative view of protein, gene and 3D structural information. *Nucleic Acids Res.* **2017,** *45* (D1), D271-D281.

294. Berman, H.; Henrick, K.; Nakamura, H., Announcing the worldwide Protein Data Bank. *Nat. Struct. Biol.* **2003,** *10* (12), 980.