



Published in final edited form as:

J Chem Theory Comput. 2017 September 12; 13(9): 4584–4592. doi:10.1021/acs.jctc.7b00500.

POVME 3.0: Software for Mapping Binding Pocket Flexibility

Jeffrey R. Wagner¹, Jesper Sørensen¹, Nathan Hensley¹, Celia Wong¹, Clare Zhu¹, Taylor Perison¹, and Rommie E. Amaro^{1,2}

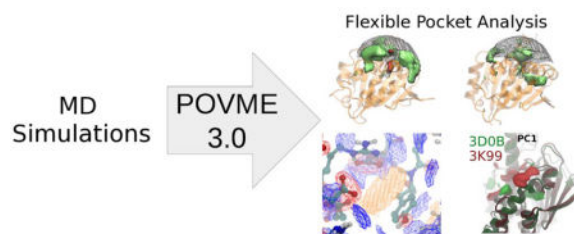
¹Department of Chemistry and Biochemistry; University of California, San Diego

²National 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.

Graphical Abstract



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.¹ 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.

Corresponding Author: Rommie E. Amaro, Ph.D., ramaro@ucsd.edu, +1 (858) 534-9629. Jeffrey R. Wagner j5wagner@ucsd.edu, 9500 Gilman Dr., Urey Hall 3234, La Jolla, CA. 92093.

Supporting Information

The following files are provided free of charge:

User notes and best practices, the 20 simulated HSP90 ligand structures, detailed clustering and PCA results, protein preparation commands, and AMBER MD input files. (.docx)

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.² While algorithms have been developed to efficiently sample a small molecule ligand's conformational landscape,^{3,4,5,6,7,8,9,10} proteins are considerably more complex and proper handling of their flexibility is a more challenging question.^{11,12,13}

The “ensemble docking” method is used to bridge the gap between the available rigid docking techniques and existing models of protein flexibility.^{14,15,16} 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, it is 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.¹⁷ 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.^{16,18,19} 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.^{19,20} 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.²¹

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.²² 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,²³ 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.²⁴

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.²⁵ Some algorithms terminate the pocket when new points are no longer adjacent to a ligand or pocket-lining residue atom.²⁴ 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.^{26,27,28} Another option, employed in POVME²⁹ and other programs^{30,31} 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^{32,29,33}, TRAPP^{24,22}, PocketAnalyzer(PCA)³⁴, trj_cavity²⁷, Epock³⁵, and Volsite.²⁶

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)³⁶ 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.²⁹ 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.³⁵ 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,³⁷ 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.³⁸³⁹ 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 penalty⁴⁰ 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.²⁸²⁴ 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 VMD⁴¹ 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).⁴² These PDB codes were selected on the basis of ligand diversity, structure resolution, and pocket characteristics. The final 20 PDB codes selected are 1BYQ,⁴³ 1UYF,⁴⁴ 1UYI,⁴⁴ 1UYL,⁴⁴ 2VCI,⁴⁵ 2WI7,⁴⁶ 2XHR,⁴⁷ 2YEJ,⁴⁸ 3B26,⁴⁹ 3D0B,⁵⁰ 3HEK,⁵¹ 3K98,⁵² 3K99,⁵² 3RKZ,⁵³ 3RLR,⁵⁴ 4CWN,⁵⁵ 4FCR,⁵⁶ 4LWE,⁵⁷ 4R3M,⁵⁸ 4W7T.⁵⁹ Active site ligands were parameterized using GAFF,⁶⁰ with charges derived using Gaussian⁶¹ and the RED server.^{62,63,64} 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.³ The FF99SB force field was

used for simulation.⁶⁵³ The `tleap solvatebox` command was used to add a TIP4P water box with 10Å 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 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 `numpy` files for visualization and custom user analysis.

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 20×20 similarity matrices. These matrices show the POVME similarity of the pockets from the first frame of each simulation (Figure 2A), the POVME similarity of the pockets from the last frame of each simulation (Figure 2B), and the average POVME similarity of all 250 frames taken from each simulation (Figure 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 20×20 ligand similarity matrix (Figure 2d).

To compare the information contained in each similarity matrix, the Kendall rank correlation coefficient⁶⁶ 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 2D), the bright red (1UYI,1UYF) hotspot has the highest nonidentical similarity value (See figure S1 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 2C) to the ligand similarity matrix (Figure 2B) yields a Kendall Tau value of 0.266, with a p of 4.90×10^{-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 5000×5000 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 \times 20 simulations). The workflow is capable of choosing a number of clusters automatically using the Kelley penalty method.⁴⁰ 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 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)⁶⁷, 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 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 S2). 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 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.

Principal Component Analysis

Principal Component Analysis was performed on the 20 HSP90 trajectories. Figure S3 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 4B).

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.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

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.

References

1. Brooijmans N, Kuntz ID. Molecular Recognition and Docking Algorithms. *Annu Rev Biophys Biomol Struct.* 2003; 32(1):335–373. [PubMed: 12574069]
2. Antunes DA, Devaurs D, Kavraki LE. Understanding the Challenges of Protein Flexibility in Drug Design. *Expert Opin Drug Discov.* 2015; 10(12):1301–1313. [PubMed: 26414598]
3. Wang J, Wang W, Kollman PA, Case DA. Automatic Atom Type and Bond Type Perception in Molecular Mechanical Calculations. *J Mol Graph Model.* 2006; 25(2):247–260. [PubMed: 16458552]
4. Betz RM, Walker RC. Paramfit: Automated Optimization of Force Field Parameters for Molecular Dynamics Simulations. *J Comput Chem.* 2015; 36(2):79–87. [PubMed: 25413259]
5. Vanommeslaeghe K, Yang M, MacKerell AD Jr. Robustness in the Fitting of Molecular Mechanics Parameters. *J Comput Chem.* 2015; 36(14):1083. [PubMed: 25826578]
6. Harder E, Damm W, Maple J, Wu C, Reboul M, Xiang JY, Wang L, Lupyan D, Dahlgren MK, Knight JL, Kaus JW, Cerutti DS, Krilov G, Jorgensen WL, Abel R, Friesner RA. OPLS3: A Force Field Providing Broad Coverage of Drug-like Small Molecules and Proteins. *J Chem Theory Comput.* 2016; 12(1):281–296. [PubMed: 26584231]
7. Vanommeslaeghe K, Hatcher E, Acharya C, Kundu S, Zhong S, Shim J, Darian E, Guvench O, Lopes P, Vorobyov I, Mackerell AD. 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.
8. Malde AK, Zuo L, Breeze M, Stroet M, Poger D, Nair PC, Oostenbrink C, Mark AE. An Automated Force Field Topology Builder (ATB) and Repository: Version 1.0. *J Chem Theory Comput.* 2011; 7(12):4026–4037. [PubMed: 26598349]
9. Ebejer J-P, Morris GM, Deane CM. Freely Available Conformer Generation Methods: How Good Are They? *J Chem Inf Model.* 2012; 52(5):1146–1158. [PubMed: 22482737]
10. Hawkins PCD, Skillman AG, Warren GL, Ellingson BA, Stahl MT. 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. [PubMed: 20235588]
11. Lexa KW, Carlson HA. Protein Flexibility in Docking and Surface Mapping. *Q Rev Biophys.* 2012; 45(3):301–343. [PubMed: 22569329]
12. Lin J-H. Accommodating Protein Flexibility for Structure-Based Drug Design. *Curr Top Med Chem.* 2011; 11(2):171–178. [PubMed: 20939792]
13. Spyarakis F, BidonChanal A, Barril X, Luque FJ. Protein Flexibility and Ligand Recognition: Challenges for Molecular Modeling. *Curr Top Med Chem.* 2011; 11(2):192–210. [PubMed: 20939788]
14. Ellingson SR, Miao Y, Baudry J, Smith JC. Multi-Conformer Ensemble Docking to Difficult Protein Targets. *J Phys Chem B.* 2015; 119(3):1026–1034. [PubMed: 25198248]
15. Sørensen J, Demir Ö, Swift RV, Feher VA, Amaro RE. Molecular Docking to Flexible Targets. *Methods in Molecular Biology.* 2014:445–469.
16. Huang S-Y, Zou X. Ensemble Docking of Multiple Protein Structures: Considering Protein Structural Variations in Molecular Docking. *Proteins.* 2007; 66(2):399–421. [PubMed: 17096427]
17. Totrov M, Abagyan R. Flexible Ligand Docking to Multiple Receptor Conformations: A Practical Alternative. *Curr Opin Struct Biol.* 2008; 18(2):178–184. [PubMed: 18302984]
18. Sørensen J, Demir Ö, Swift RV, Feher VA, Amaro RE. Molecular Docking to Flexible Targets. *Methods in Molecular Biology.* 2014:445–469.
19. Wong CF. Flexible Receptor Docking for Drug Discovery. *Expert Opin Drug Discov.* 2015; 10(11):1189–1200. [PubMed: 26313123]

20. Osguthorpe DJ, Sherman W, Hagler AT. Exploring Protein Flexibility: Incorporating Structural Ensembles from Crystal Structures and Simulation into Virtual Screening Protocols. *J Phys Chem B*. 2012; 116(23):6952–6959. [PubMed: 22424156]
21. Henrich S, Salo-Ahen OMH, Huang B, Rippmann FF, Cruciani G, Wade RC. Computational Approaches to Identifying and Characterizing Protein Binding Sites for Ligand Design. *J Mol Recognit*. 2010; 23(2):209–219. [PubMed: 19746440]
22. Kokh DB, Czodrowski P, Rippmann F, Wade RC. Perturbation Approaches for Exploring Protein Binding Site Flexibility to Predict Transient Binding Pockets. *J Chem Theory Comput*. 2016; 12(8):4100–4113. [PubMed: 27399277]
23. Stank A, Kokh DB, Fuller JC, Wade RC. Protein Binding Pocket Dynamics. *Acc Chem Res*. 2016; 49(5):809–815. [PubMed: 27110726]
24. Kokh DB, Richter S, Henrich S, Czodrowski P, Rippmann F, Wade RC. TRAPP: A Tool for Analysis of Transient Binding Pockets in Proteins. *J Chem Inf Model*. 2013; 53(5):1235–1252. [PubMed: 23621586]
25. Kleywegt GJ, Jones TA. Detection, Delineation, Measurement and Display of Cavities in Macromolecular Structures. *Acta Crystallogr D Biol Crystallogr*. 1994; 50(Pt 2):178–185. [PubMed: 15299456]
26. 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. [PubMed: 22834646]
27. Paramo T, East A, Garzón D, Ulmschneider MB, Bond PJ. Efficient Characterization of Protein Cavities within Molecular Simulation Trajectories: Trj_cavity. *J Chem Theory Comput*. 2014; 10(5):2151–2164. [PubMed: 26580540]
28. Craig IR, Pflieger C, Gohlke H, Essex JW, Spiegel K. Pocket-Space Maps to Identify Novel Binding-Site Conformations in Proteins. *J Chem Inf Model*. 2011; 51(10):2666–2679. [PubMed: 21910474]
29. Durrant JD, Votapka L, Sørensen J, Amaro RE. POVME 2.0: An Enhanced Tool for Determining Pocket Shape and Volume Characteristics. *J Chem Theory Comput*. 2014; 10(11):5047–5056. [PubMed: 25400521]
30. 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. [PubMed: 9761470]
31. Saberi Fathi S, Fathi SS, Tuszynski JA. A Simple Method for Finding a Protein’s Ligand-Binding Pockets. *BMC Struct Biol*. 2014; 14(1):18. [PubMed: 25038637]
32. Le Guilloux V, Schmidtke P, Tuffery P. Fpocket: An Open Source Platform for Ligand Pocket Detection. *BMC Bioinformatics*. 2009; 10:168. [PubMed: 19486540]
33. 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–W589. [PubMed: 20478829]
34. Craig IR, Pflieger C, Gohlke H, Essex JW, Spiegel K. Pocket-Space Maps to Identify Novel Binding-Site Conformations in Proteins. *J Chem Inf Model*. 2011; 51(10):2666–2679. [PubMed: 21910474]
35. Laurent B, Chavent M, Cragolini T, Dahl ACE, Pasquali S, Derreumaux P, Sansom MSP, Baaden M. Epocket: Rapid Analysis of Protein Pocket Dynamics. *Bioinformatics*. 2015; 31(9):1478–1480. [PubMed: 25505095]
36. Durrant JD, de Oliveira CAF, McCammon JA. POVME: An Algorithm for Measuring Binding-Pocket Volumes. *J Mol Graph Model*. 2011; 29(5):773–776. [PubMed: 21147010]
37. Durrant JD, McCammon JA. BINANA: A Novel Algorithm for Ligand-Binding Characterization. *J Mol Graph Model*. 2011; 29(6):888–893. [PubMed: 21310640]
38. Oliphant TE. Python for Scientific Computing. *Comput Sci Eng*. 2007; 9(3):10–20.
39. Millman KJ, Jarrod Millman K, Aivazis M. Python for Scientists and Engineers. *Comput Sci Eng*. 2011; 13(2):9–12.

40. Kelley LA, Gardner SP, Sutcliffe MJ. An Automated Approach for Clustering an Ensemble of NMR-Derived Protein Structures into Conformationally Related Subfamilies. *Protein Eng.* 1996; 9(11):1063–1065. [PubMed: 8961360]
41. Humphrey W, Dalke A, Schulten K. VMD: Visual Molecular Dynamics. *J Mol Graph.* 1996; 14(1): 33–38. [PubMed: 8744570]
42. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. *Nucleic Acids Res.* 2000; 28(1):235–242. [PubMed: 10592235]
43. Obermann WM, Sondermann H, Russo AA, Pavletich NP, Hartl FU. In Vivo Function of Hsp90 Is Dependent on ATP Binding and ATP Hydrolysis. *J Cell Biol.* 1998; 143(4):901–910. [PubMed: 9817749]
44. 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 RE. Structure-Activity Relationships in Purine-Based Inhibitor Binding to HSP90 Isoforms. *Chem Biol.* 2004; 11(6):775–785. [PubMed: 15217611]
45. Brough PA, Aherne W, Barril X, Borgognoni J, Boxall K, Cansfield JE, Cheung K-MJ, Collins I, Davies NGM, Drysdale MJ, Dymock B, Eccles SA, Finch H, Fink A, Hayes A, Howes R, Hubbard RE, James K, Jordan AM, Lockie A, Martins V, Massey A, Matthews TP, McDonald E, Northfield CJ, Pearl LH, Prodromou C, Ray S, Raynaud FI, Roughley SD, Sharp SY, Surgenor A, Walmsley DL, Webb P, Wood M, Workman P, Wright L. 4,5-Diarylisoaxazole Hsp90 Chaperone Inhibitors: Potential Therapeutic Agents for the Treatment of Cancer. *J Med Chem.* 2008; 51(2): 196–218. [PubMed: 18020435]
46. Brough PA, Barril X, Borgognoni J, Chene P, Davies NGM, Davis B, Drysdale MJ, Dymock B, Eccles SA, Garcia-Echeverria C, Fromont C, Hayes A, Hubbard RE, Jordan AM, Jensen MR, Massey A, Merrett A, Padfield A, Parsons R, Radimerski T, Raynaud FI, Robertson A, Roughley SD, Schoepfer J, Simmonite H, Sharp SY, 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. [PubMed: 19610616]
47. Murray CW, Carr MG, Callaghan O, Chessari G, Congreve M, Cowan S, Coyle JE, Downham R, Figueroa E, Frederickson M, Graham B, McMenemy R, O'Brien MA, Patel S, Phillips TR, Williams G, Woodhead AJ, Woolford AJ-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. [PubMed: 20718493]
48. Roughley SD, Hubbard RE. How Well Can Fragments Explore Accessed Chemical Space? A Case Study from Heat Shock Protein 90. *J Med Chem.* 2011; 54(12):3989–4005. [PubMed: 21561141]
49. Miura T, Fukami TA, 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. [PubMed: 21875802]
50. Barta TE, Veal JM, Rice JW, Partridge JM, Fadden RP, Ma W, Jenks M, Geng L, Hanson GJ, Huang KH, Barabasz AF, Foley BE, Otto J, Hall SE. Discovery of Benzamide Tetrahydro-4H-Carbazol-4-Ones as Novel Small Molecule Inhibitors of Hsp90. *Bioorg Med Chem Lett.* 2008; 18(12):3517–3521. [PubMed: 18511277]
51. Cho-Schultz S, Patten MJ, Huang B, Elleraas J, Gajiwala KS, Hickey MJ, Wang J, Mehta PP, Kang P, Gehring MR, Kung P-P, Sutton SC. Solution-Phase Parallel Synthesis of Hsp90 Inhibitors. *J Comb Chem.* 2009; 11(5):860–874. [PubMed: 19583220]
52. Kung P-P, Huang B, Zhang G, Zhou JZ, Wang J, Digits JA, Skaptason J, Yamazaki S, Neul D, Zientek M, Elleraas J, Mehta P, Yin M-J, Hickey MJ, Gajiwala KS, Rodgers C, Davies JF, Gehring MR. Dihydroxyphenylisoindoline Amides as Orally Bioavailable Inhibitors of the Heat Shock Protein 90 (hsp90) Molecular Chaperone. *J Med Chem.* 2010; 53(1):499–503. [PubMed: 19908836]
53. Zapf CW, Bloom JD, Li Z, Dushin RG, Nittoli T, Otteng M, Nikitenko A, Golas JM, Liu H, Lucas J, Boschelli F, Vogan E, Olland A, Johnson M, Levin JI. 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. [PubMed: 21715165]

54. Kung P-P, Sinnema P-J, Richardson P, Hickey MJ, Gajiwala KS, Wang F, Huang B, McClellan G, Wang J, Maegley K, Bergqvist S, Mehta PP, Kania R. Design Strategies to Target Crystallographic Waters Applied to the Hsp90 Molecular Chaperone. *Bioorg Med Chem Lett*. 2011; 21(12):3557–3562. [PubMed: 21612924]
55. Casale E, Amboldi N, Brasca MG, Caronni D, Colombo N, Dalvit C, Felder ER, 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. [PubMed: 24980703]
56. Davies NGM, Browne H, Davis B, Drysdale MJ, Foloppe N, Geoffrey S, Gibbons B, Hart T, Hubbard R, Jensen MR, Mansell H, Massey A, Matassova N, Moore JD, Murray J, Pratt R, Ray S, Robertson A, Roughley SD, 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. [PubMed: 23018093]
57. 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. [PubMed: 25313505]
58. 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. [PubMed: 25490978]
59. McBride CM, 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 TE, Abrams T, Fong S, Hendrickson T, Lei D, Lin J, Menezes D, Pryer N, Taverna P, Xu Y, Zhou Y, Shafer CM. Design, Structure-Activity Relationship, and in Vivo Characterization of the Development Candidate NVP-HSP990. *J Med Chem*. 2014; 57(21):9124–9129. [PubMed: 25368984]
60. Wang J, Wolf RM, Caldwell JW, Kollman PA, Case DA. Development and Testing of a General Amber Force Field. *J Comput Chem*. 2004; 25(9):1157–1174. [PubMed: 15116359]
61. Gaussian 09. Gaussian, Inc; Wallingford CT: 2009.
62. Vanqualef E, Simon S, Marquant G, Garcia E, Klimerak G, Delepine JC, Cieplak P, Dupradeau F-YRED. 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. [PubMed: 21609950]
63. Dupradeau F-Y, Pigache A, Zaffran T, Savineau C, Lelong R, Grivel N, Lelong D, Rosanski W, Cieplak P. The RED Tools: Advances in RESP and ESP Charge Derivation and Force Field Library Building. *Phys Chem Chem Phys*. 2010; 12(28):7821. [PubMed: 20574571]
64. Bayly CI, Cieplak P, Cornell W, Kollman PA. 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.
65. 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. [PubMed: 16981200]
66. Kendall MG. A NEW MEASURE OF RANK CORRELATION. *Biometrika*. 1938; 30(1–2):81–93.
67. Vauquelin G, Charlton SJ. Long-Lasting Target Binding and Rebinding as Mechanisms to Prolong in Vivo Drug Action. *Br J Pharmacol*. 2010; 161(3):488–508. [PubMed: 20880390]

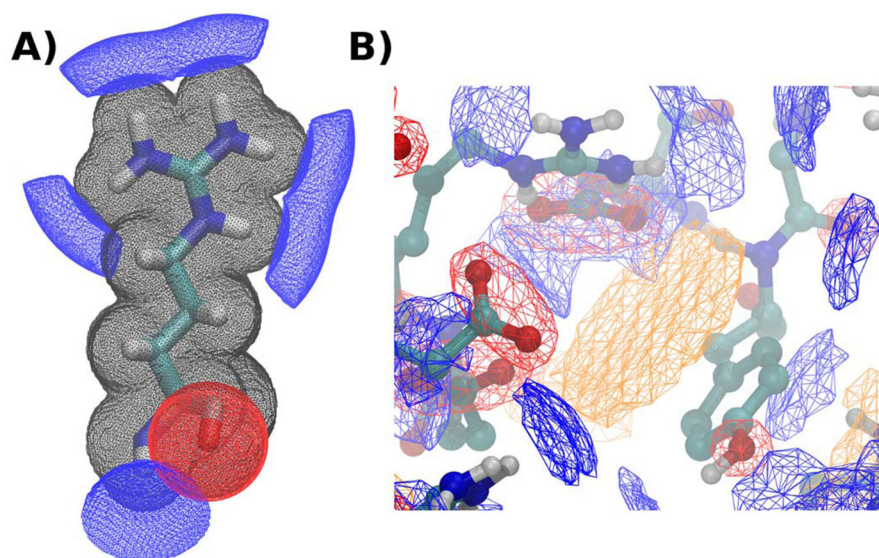


Figure 1. POVME coloring scheme. Black mesh shows occupancy, blue shows hydrogen bond donor regions, red shows hydrogen bond acceptor, and orange shows pi-stacking. A) The POVME coloring scheme applied at 0.1 Angstrom grid resolution to a single arginine residue from a protein structure. B) The POVME coloring scheme applied to a binding pocket at 0.75 Angstrom resolution, with occupancy not shown.

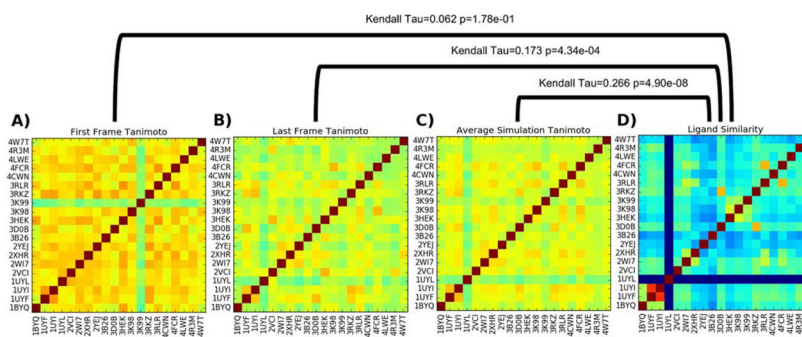
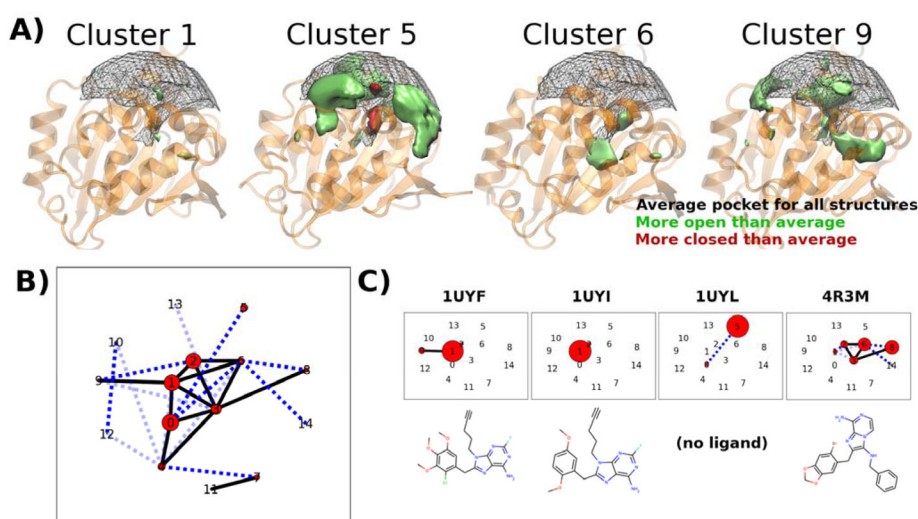


Figure 2. Tanimoto similarity matrices of A) initial simulation pocket shapes (following equilibration), B) final simulation pocket shapes (after 250 ns MD), C) average similarity of all pocket shapes throughout simulation, and D) RDKit chemical fingerprint Tanimoto score. Kendall Tau analyses of the matrices are shown above the brackets.

**Figure 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.

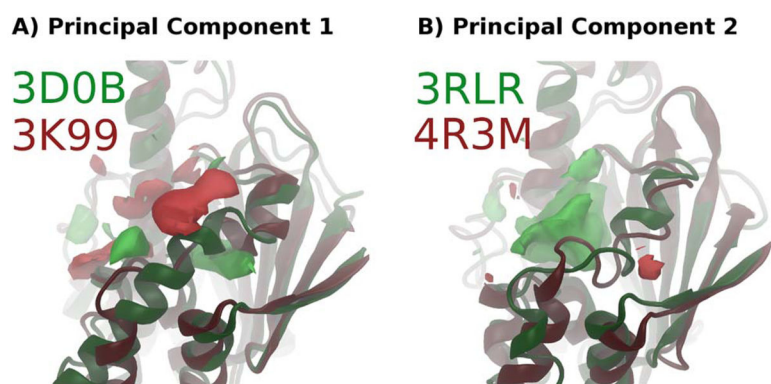


Figure 4.

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.