Eurographics Workshop on Visual Computing for Biology and Medicine (2019)
B. Kozlíková, L. Linsen, and P.-P. Vázquez (Editors)

DockVis: Visual Analysis of Molecular Docking Data

K. Furmanová[1], B. Kozlíková[1], V. Vonásek[2], and J. Byška[1,3]

[1]Masaryk University, Brno, Czech Republic
[2]Czech Technical University in Prague, Prague, Czech Republic
[3]University of Bergen, Norway

Abstract

Molecular docking is one of the key mechanisms for predicting possible interactions between ligands and proteins. This highly complex task can be simulated by several software tools, providing the biochemists with possible ligand trajectories, which have to be subsequently explored and evaluated for their biochemical relevance. This paper focuses on aiding this exploration process by introducing DockVis visual analysis tool. DockVis operates primarily with the multivariate output data from one of the latest available tools for molecular docking, CaverDock. CaverDock output consists of several parameters and properties, which have to be subsequently studied and understood. DockVis was designed in tight collaboration with protein engineers using the CaverDock tool. However, we believe that the concept of DockVis can be extended to any other molecular docking tool providing the users with corresponding computation results.

CCS Concepts

• ***Human-centered computing*** → *Scientific visualization; Visualization systems and tools;*

1. Introduction

Protein structure and function can be influenced and changed through interactions with other molecules. In protein engineering and drug design, the crucial interactions are those between the protein and a small ligand molecule. Ligands are usually entering the protein inner structure and performing a chemical reaction in a deeply buried active site. The simulation of transportation of the ligand to the active site is the main task for molecular docking algorithms. This very challenging task has been already investigated by many research groups. As a result, there are several tools enabling to calculate possible ligand dockings [MLW08]. The naïve approach is searching a high-dimensional space and trying to detect all possible entrance paths for ligand transportation to the active site and calculating their geometric and physico-chemical properties. Without any further optimization and guidance, this process can be very lengthy and based on the quality of the scoring functions of the respective tools, the resulting dataset of possible paths can be vast.

As the problem of molecular docking is still highly challenging, new methods for its calculation are appearing. Recently, one of the most promising methods for "guided" molecular docking was released by Filipovič et al. [FVP*19]. Their CaverDock tool uses the knowledge about a possible ligand entrance path, called tunnel. Such a tunnel is computed purely geometrically, using one of the existing and widely adopted tools, such as CAVER [CPB*12] or Mole [SSVB*13]. An overview of the existing algorithms and methods for tunnel calculation, along with their visual representation, can be found in the survey by Krone et al. [KKL*16].

CaverDock utilizes the detected tunnel for navigating the ligand to the protein active site and calculates the possible passage of the ligand. The algorithm outputs a sequence of consecutive ligand configurations and corresponding energy profiles. These need to be further analyzed by biochemists and the most biochemically relevant dockings need to be tested in a laboratory.

The analysis of data produced by the docking systems is a challenging task as the computational tools usually provide no or only limited visual support. The biochemists need to combine many geometric and physico-chemical properties of the computed ligand docking to assess its feasibility. This includes studying the ligand conformation changes within the transportation, energy profiles, interactions of the ligand with the surrounding amino acids, properties of amino acids, and others. The existing approaches allow to visualize mostly only the 3D representation of protein and ligand and animate the calculated docking. However, the rest of the crucial properties is completely omitted. Therefore, the biochemists are forced to use several independent tools to create visual representations of these properties and manually combine the information, without any option for interactivity and interplay between them.

Therefore, in this paper we focus on this problem and propose a novel visual analysis tool, DockVis, enabling the user to load the results of the CaverDock tool and explore many properties at once and in an interactive manner. To reach that, we integrate several linked views, both spatial and abstracted, aiming to provide the

DOI: 10.2312/vcbm.20191238

biochemists with all-in-one solution for exploration of the transportation of a ligand to the active site.

The proposed tool has been designed in tight collaboration with protein engineers to address their urgent needs. Although the tool has been tailored to interplay with the results obtained by the CaverDock tool, we believe that the principles can be applied to any of the other computational tools that produce trajectories passing through the same protein tunnel.

In summary, the main contributions of this paper are:

- Analysis of requirements for visual investigation and comparison of two spatially related ligand trajectories.
- The means to spatially and temporally align these two trajectories and compare them based on multiple properties.
- The means to detect and visualize conformational changes of the ligand along the trajectories.
- The means to relate the ligand spatial conformation with its surroundings in an abstracted view.

2. Related Work

In this section, we will describe the most important approaches for molecular docking calculation as well as for the visualization of ligand passage through a protein tunnel.

2.1. Molecular Docking

Molecular docking aims to predict the most probable binding of two molecules; in our case a ligand to a given protein. It requires to examine many mutual conformations of the ligand and protein. The quality of the binding can be evaluated using the potential energy of the system, which is usually approximated by a force-field (used, e.g., in early Autodock3 [MGH*98]), or using empirical scoring functions. The scoring functions, besides the energy term, also consider other properties, such as electrostatics, entropy, hydrophobicity, etc. Examples of the scoring functions are ChemScore [EMA*97], PLANTS [KSE09], or Autodock4 [HMOG07]. Early methods evaluated the binding based on the geometric fit of the molecules [Mez03].

Searching for the best ligand pose depends on the overall flexibility of the system. The early methods considered both the ligand and protein to be rigid [Mez03]. This leads to search in 6D space (i.e, position and rotation), where the binding is evaluated based on the shape complementarity using Fast Fourier Transform (FFT) [KKSE*92]. FFT-based tools are, for example, ZDOCK [BFR00] and MEGADOCK [OSS*14].

However, experiments show that ligand and protein should be considered as flexible [Ham02], which motivates the development of new tools. In semi-flexible docking, only one of the molecules is considered to be flexible; the second molecule remains rigid. This leads to search in the (6+N)dimensional space (with N conformational variables), which requires a different approach. Among the most notable methods belong the systematic and randomized search. The systematic search requires to strictly decrease the number of rotations to limit the search space [FBM*04]. Ligand's flexibility can be also emulated using a set of conformations [SLVM08]. The randomized methods are more suitable for searching the high-dimensional spaces. For example, in Monte-Carlo-based search, actual conformation is changed randomly in each iteration. This change is accepted if the new conformation has a better score. Otherwise, the new configuration is accepted with the probability determined by the difference between the two scores. Tools using this search are, e.g., MCDOCK [LW99], AutoDock Vina [TO10], and RosettaLigand [MB06]. Finally, sampling-based motion planning approaches, originally developed in robotics [ABSC12], were applied to conformational search, e.g., in the MoMa-LigPath tool [CLIS10].

Considering the full flexibility of both protein and ligand significantly increases the dimension of the space to be searched. A possible approach to decrease the search space is docking a flexible ligand to several static protein conformations. These static protein conformations can be obtained from Molecular Dynamics (MD) that computes the behavior of a protein without any ligand [ABM08].

However, predicting the possible binding of molecules with MD simulations is very computationally demanding. Therefore, most of the described tools are focusing on the final binding while ignoring the access pathway. The very recently released tool, CaverDock [VFP*19, FVP*19], tries to overcome this issue by combining the docking and pathway traversability. Overall, molecular docking is a very active research area with many designed tools so we refer to recent surveys for more details [SM18, PST17].

2.2. Visualization of Ligand Trajectory

Visualization and visual analysis of ligand passage through a protein have been already in focus of several research groups. The simplest visual representation of ligand transportation can be reached by animating the movement of the ligand in a frame-by-frame manner using one of the state-of-the-art general-purpose molecular visualization tools, such as PyMol [Sch15] or VMD [HDS96].

Except for that, there are several specialized tools and visualizations, enabling to investigate several additional features of the ligand transportation. Furmanová et al. [FJB*17] proposed a visual analysis tool for exploration of several geometric properties of the ligand passage, such as the free space around the ligand or its speed. Additionally, they proposed methods for smoothing the originally very scattered ligand trajectory to be able to reveal trends in ligand movement. However, this tool cannot be used for comparison of multiple trajectories and it does not operate with chemical properties, such as the binding energy.

The ligand binding energy was studied in a very recent approach by Jurčík et al. [JFB*19]. In their case, they analyzed large ensembles of ligand trajectories and their energy profiles. These trajectories were calculated using a path planning method which does not provide the users with the information about the tunnel. Therefore, it is also not directly applicable to our case.

Duran et al. [DHR*19] presented another tool for interactive analysis of ligand trajectories, combining 3D view with 2D plots of several properties (both geometric and chemical). Their tool focuses on suggesting the potentially interesting parts of the molecular dynamics simulation to the user by highlighting the background of the graphs. A lot of effort is given to interaction and linking between the views. The tool also enables to compare the trajectories of few ligands entering the same protein structure. However, it does not operate with tunnels and does not support the interplay between data obtained by molecular docking.

Another approach was proposed by Hermosilla et al. [HEG*17]. Their tool enables to visualize the interaction forces between ligand and the protein amino acids. Although the task is very similar, the tool does not provide the users with the energy plots and the information about the tunnel.

There are also several approaches operating with abstracted representations of ligand and its interactions with the protein. LigPlot+ [LS11] is a typical representative of these approaches. This tool automatically transforms the 3D representation to a 2D diagram of a ligand and amino acids of the interacting protein which are in the reaction distance to the ligand. The diagram then depicts the interactions as dashed lines between the corresponding atoms. More recently, Vázquez et al. [VHG*18] published their system for compact 2D visualization of molecular dynamics simulations, capturing the protein-ligand interaction. The central part of their proposed visualization shows the ligand, while the protein is encoded into a circular layout around the ligand. Although the abstracted views are already well adopted by biochemists and for small molecular structures they give a very comprehensible overview of the structure, these cannot be used in our case without any supplementary view, showing the other molecular docking properties. However, in our solution we were inspired by the concept of the mentioned abstracted views and they form an important part of our proposed visual analysis tool.

3. Data and Tasks Abstractions

To understand the design rationale of our proposed tool, we have to first describe the input data and the tasks which were extracted from numerous discussions with the biochemists. The input data are coming from the CaverDock tool. Therefore, in the following section this tool will be briefly introduced.

3.1. CaverDock

CaverDock [VFP*19, FVP*19] simulates the transportation of a ligand through a tunnel using a step-wise algorithm. First, the input spheres representing the tunnel are sliced into a sequence of discs (see Figure 1). Second, the ligand is fitted into each disc separately such that one atom of the ligand, called a drag atom, which is selected prior to the simulation, has to be always placed within the tunnel disc corresponding to a given frame of the simulation. The best orientation of the ligand is then selected, based on the scoring function from AutoDock Vina [TO10]. Finally, the algorithm produces two distinct ligand trajectories, called lower-bound and upper-bound. The trajectory is represented as a sequence of frames, where each frame contains the positions of all atoms of the ligand.

The lower-bound trajectory is formed by gathering all ligand conformations from the successive discs, irrespective of their mutual orientation. It provides the most energetically favorable transportation of the ligand through the tunnel, but it can contain fast (non-continuous) changes of the ligand orientation.

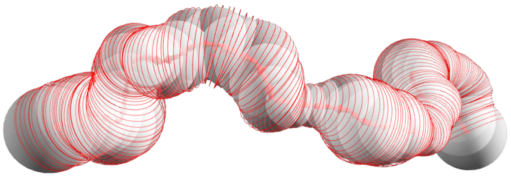


Figure 1: *Illustration of the tunnel sliced into a set of discs (red circles), used in CaverDock. The tunnel itself is represented by a set of gray spheres and its centerline is denoted by red arrows. Image taken from [FVP*19].*

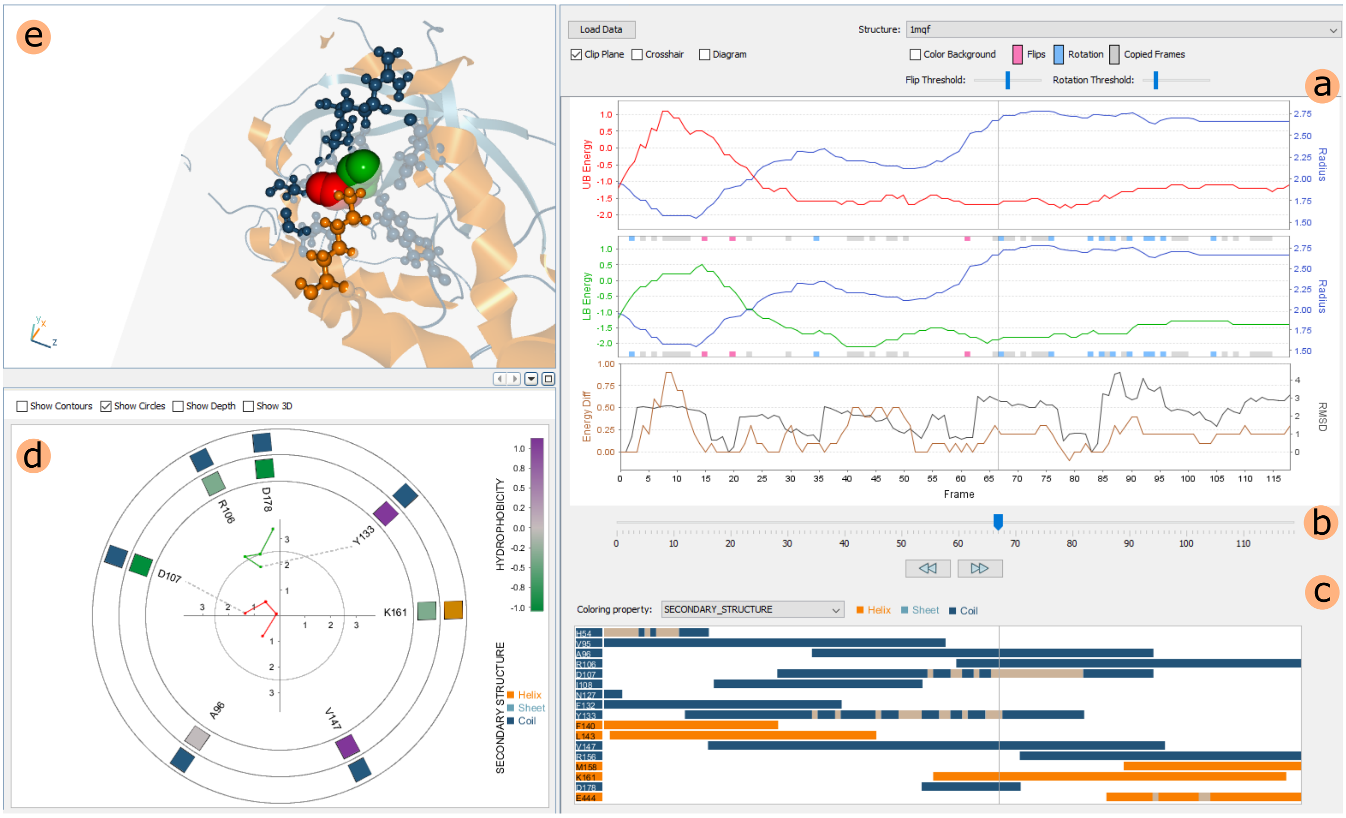
On the other hand, when computing the upper-bound trajectory, the algorithm considers the position of the ligand in the previous disc already during the fitting process and allows only small changes of its conformation. To ensure that the final path is energetically favorable, the algorithm compares the current binding free energy in each step with the lower-bound trajectory. If the energy in lower-bound and upper-bound is too different, the algorithm inserts several frames into the upper-bound trajectory. Within these frames, a smooth transformation (e.g., rotation) between two ligand conformations is performed such that the current energy is lowered. This step aims to remove the problem of non-continuous ligand orientations in the neighboring frames of the lower-bound trajectory. These artificial frames break the correspondence between the lower-bound and upper-bound trajectories which complicates their frame-by-frame comparison.

In summary, the input data to our workflow consists of protein structure, two trajectories, and the sequence of discs describing the tunnel used by both of the trajectories. Each disc is defined by its center, radius, and normal vector determining its orientation. The trajectories are formed by sequence of frames describing consequent ligand conformations. Each frame corresponds to one of the discs. The trajectories can be of different length but they do not contain backward movements.

3.2. Tasks

When analyzing the CaverDock results, the experts have to evaluate both lower-bound and upper-bound trajectories as they provide the complementary information. The lower-bound trajectory gives information about the best possible scenario concerning energetic efficiency. As such, it can be used to compare multiple datasets. However, it also often ignores smaller energetic bottlenecks which can be skipped due to the rapid changes of the ligand conformations. On the other hand, the upper-bound trajectory is continuous and smooth but may not be energetically optimal as the algorithm can generate very unfavorable conformations in the narrow parts of the tunnel.

When the domain experts are investigating the lower- and upper-bound trajectories, they are mostly interested in the exploration of their energy profiles. These profiles provide the information about the likelihood of a particular ligand-protein conformation happening in real world or laboratory — the higher energy values are an indicative of energetic barriers (i.e., places requiring more energy than available) that the ligand may not be able to pass. When the



Load Data

Structure: 1mqf

☑ Clip Plane □ Crosshair □ Diagram

□ Color Background Flips Rotation Copied Frames

Flip Threshold: Rotation Threshold:

Coloring property: SECONDARY_STRUCTURE

□ Show Contours ☑ Show Circles □ Show Depth □ Show 3D

Figure 2: *Overview of our system: a) Energy Profiles, b) Navigation Slider c) Lining Residues Plot, d) Ligand View, and e) 3D View where the protein is displayed using cartoon representation, with α-helices depicted by orange and β-sheets by light blue color.*

energetic barrier is too high, the upper-bound trajectory may not be calculated at all.

However, the ligand binding energy itself does not provide enough information for some tasks. For instance, in protein engineering, one of the ultimate goals is to strengthen or block particular protein-ligand interactions. Here the domain experts are either designing new regulatory molecules (i.e., activators or inhibitors) that regulate the speed of reaction or they directly modify the protein itself to achieve their goals.

When developing new activators and inhibitors, the investigation of possible hydrogen bonds between the new regulatory molecule and protein plays a crucial role as their presence can suggest more stable binding. On the other hand, when modifying the protein itself, the task is to identify particular amino acids that should be mutated. Here the protein engineers need to relate the energy profiles to the protein structure as the energy may be used to identify the problematic parts that need to be mutated. Namely, the geometry of the protein tunnel and the physico-chemical properties of amino acids in the tunnel vicinity are the most important parameters.

Finally, when investigating the lower-bound trajectory, it is important to detect sudden changes in ligand conformation and explore them. Such changes can be either artifacts caused by the missing continuity constraint or indicate natural flexibility of the protein which is not possible to simulate in CaverDock at the moment. Here, the knowledge about the orientation of protein secondary structures in the vicinity of the ligand can be very helpful. Secondary structures represent patterns in the protein structure, the most common types are α-helices and β-sheets (see Figure 2e). As the secondary structures are mostly very rigid formations and thus stable within the molecular dynamics, they can be used to estimate protein flexibility in a given area.

To identify all requirements for a system that would enable the visual investigation of the CaverDock results, we have conducted several informal interviews with the authors of the tool. Based on these interviews, we have identified the following set of crucial tasks that our visualization system has to support:

- **R1** Enable comparison between the upper- and lower-bound trajectories to identify important differences.
- **R2** Enable the identification of sudden changes of ligand orientation or conformation and relate them to changes in the binding energy.
- **R3** Provide an intuitive way to study the surroundings of the ligand in selected parts of its trajectory.
- **R4** Relate the ligand trajectory to the protein tunnel used by CaverDock.
- **R5** Identify types and orientation of secondary structures in the vicinity of the ligand during the transportation.
- **R6** Enable the identification of hydrogen bonds.

4. Proposed Solution

Based on the requirements, we designed our system in the following way. It consists of five main parts (see Figure 2): Energy Profiles (a) provide an initial overview and comparison of the trajectories. Navigation Slider (b) allows to select a single frame for detailed analysis. Lining Residues Plot (c) depicts the amino acids surrounding the tunnel disc at a given frame. The Energy Profiles and Lining Residues Plot are interactively linked with the Ligand View (d) that shows the ligands (its position in the lower-bound and upper-bound trajectory) and the lining amino acids projected to the plane, given by the tunnel disc corresponding to the selected frame. The same information is also depicted in the 3D View (e). In the following sections, we will describe the individual parts of our system and how they address the previously specified tasks.

4.1. Trajectory Alignment

As described in Section the 3.1, the lower-bound trajectory is formed by the most energetically favorable ligand conformations in each tunnel disc, while the upper-bound trajectory restricts the conformational and energetic changes. It often contains multiple continuously changing ligand conformations per single tunnel disc. As a result, the upper-bound trajectory is usually longer in terms of frames. To be able to compare the properties of these trajectories (**R1**), we need to align them. For each frame of the trajectory we have the information about its position within the tunnel (i.e., the tunnel disc ID). Therefore, we simultaneously iterate over the lower- and upper-bound trajectories and check the ID of the tunnel disc. In places where the upper-bound trajectory contains multiple consecutive frames at the same tunnel disc, we extend the lower-bound trajectory by inserting the copied frames corresponding to the same tunnel disc. This way we ensure that the resulting trajectories contain the same number of frames and the parallel frames correspond to the same tunnel disc.

4.2. Energy Profiles

The initial overview of the dataset is provided by the Energy Profiles chart (Figure 3). Here we depict the energy profiles of the lower- (green) and upper-bound (red) trajectories, as well as the tunnel disc radius (blue) in the corresponding trajectory frames. In most cases it can be observed that the high binding energy corresponds to the geometric bottlenecks in protein tunnels.

As we mentioned in the previous section, the lower-bound trajectory is aligned to the upper-bound trajectory by inserting the copied frames. We annotate the corresponding sections of the lower-bound energy chart with light grey segments lining the chart along the top and bottom border. These annotations serve two purposes. First, we need to make sure that the altered portions of the lower-bound trajectory are easily identifiable. Second, the alignment is performed in places where the ligand from the upper-bound trajectory lingers in the same place for multiple consecutive frames. This means that the ligand was not able to easily pass through the corresponding portion of the tunnel and the domain experts are likely to be interested in the causes of this problem.

To address the need for fast identification of geometrical changes within the trajectory (**R2**), we further annotate the frames with such changes. We distinguish between two types of geometric changes: the rotation and conformation change. We identify these changes by computing the root mean square deviation (RMSD) of the ligand atoms in two consecutive frames. If the value is above a user-defined threshold, we compute the alignment of the ligand conformations from the two frames using the Combinatorial Extension algorithm [SB98]. Then we compute the RMSD of the aligned conformations. We can then derive that if the RMSD before alignment was high, but the RMSD after alignment was equal or very close to zero, the main reason of the detected change was rotation (since the rotational changes are eliminated by the alignment). However, if the RMSD is still high after the alignment, the ligand underwent also some conformational changes. These changes are again marked as segments along the borders of the energy plots – rotation in light blue and conformation change in pink. To make the coloring more prominent, the users can choose to use full-height indicators (i.e., color the background of the whole frame) according to these changes (see Figure 7). It should be stated that we compute the geometric changes for both lower- and upper-bound trajectories. However, the restrictions imposed during the computation of the upper-bound trajectories prevent large changes of ligand conformation in the consecutive frames. Thus, unless the RMSD threshold value is set close to zero, the geometric changes are only detected in the lower-bound trajectory.

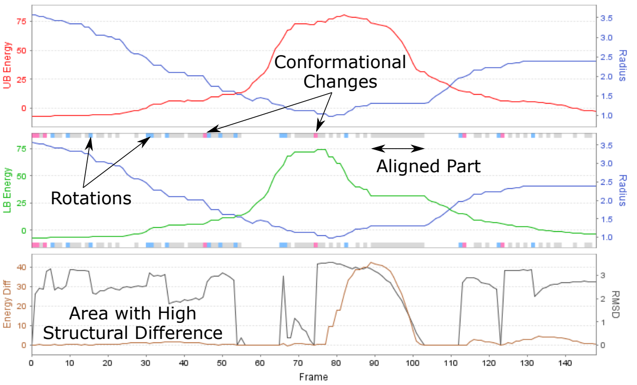


Figure 3: *Energy profiles. The first two charts show the energetic profiles of the upper-bound (red) and lower-bound (green) trajectories. The blue line indicates the width of protein tunnel. The bottom chart shows the comparison between these two trajectories. The brown line shows the difference in the binding energy while the black profile indicates the RMSD of the two ligands.*

The last chart in this view serves for direct comparison of the lower- and upper-bound trajectories (**R1**). It depicts the energetic difference of the two trajectories (brown), as well as the geometrical difference of the ligands in the corresponding frames (black). The geometrical difference is again computed as the RMSD of the ligands' atom positions. The relationship between the ligand geometry and its energetic stability is of particular interest to domain experts. Using this plot, they can easily identify portions of the trajectory where the geometric conformation of the ligand has a significant effect on the binding energy. We observed that this usually happens in narrow parts of the protein tunnel. In these parts of the trajectories, the energetic and geometric differences were often aligned and converging to zero as the ligand passed through the tunnel bottleneck. This indicates that in both cases (lower- and upper-bound) the ligand was only able to pass through this portion of the tunnel in one specific conformation. In the broader parts of the tunnel, the ligand conformations differed significantly but the spatial conformation did not affect the binding energy.

4.3. Lining Residues Plot

To enable the exploration of the amino acids surrounding the ligand during the transportation (**R3**), we provide the users with their overview within the Lining Residues Plot. Here, we first extract all amino acids that are within the interaction distance from the protein tunnel wall. We list each amino acid in a separate row and then indicate the presence of the amino acid in the individual frames by coloring the corresponding portion of the row (see Figure 4). Navigation Slider placed between the Energy Profiles and Lining Residues Plot (see Figure 2b) enables the selection of a frame which is then highlighted with a vertical line in both views. Thus, the users can easily identify amino acids surrounding the molecular tunnel in the selected frame. The frames are colored according to the selected property of the amino acids, e.g., their hydrophobicity or the type of secondary structure (**R5**). We also extract the amino acids which form the hydrogen bonds with the ligands. As the hydrogen bonds influence the transportation of the ligands, the domain experts are particularly interested where these bonds were formed and how they affect the ligand trajectory (**R6**). Therefore, we indicate the frames where the bonds were detected as light brown parts of the corresponding amino acid rows in the Lining Residues Plot. The amino acids can be selected in this plot and highlighted in 3D.

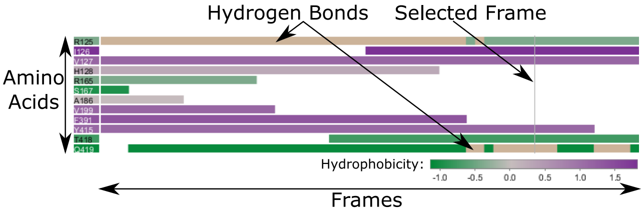


Figure 4: *The Lining Residues Plot depicting the amino acids surrounding the tunnel that are colored according to their hydrophobicity. The hydrogen bonds formed along the trajectory between the amino acids and the ligand are indicated by light brown color. The vertical line indicates the currently examined frame.*

4.4. Ligand View

To provide more detailed information about a single frame (**R3**), we adapted the MoleCollar View from [BJG*15] (see Figure 5). In this view, we display the abstracted representation of amino acids surrounding the tunnel disc at a given frame. The amino acids are depicted as bars colored by multiple user-defined properties, arranged in concentric circles (one circle per property) around the tunnel representation in the central part of the view.

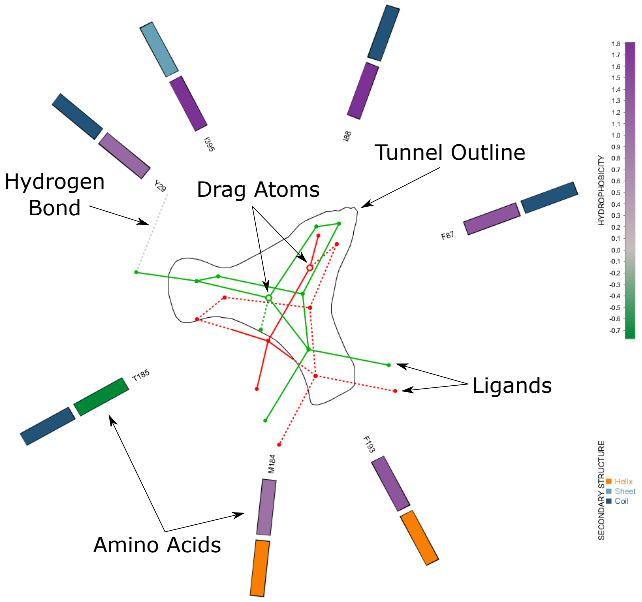


Figure 5: *Ligand View showing the abstracted representation of a single trajectory frame. The innermost contour represents the tunnel shape. The rectangles around the contour represent the amino acids surrounding the tunnel. They are colored according to their properties. The view includes ligands corresponding to lower-bound (green) and upper-bound (red) trajectories. The grey dashed line indicates the detected hydrogen bond. The orientation of the amino acids and the ligands is preserved. The part of ligand behind the disc is depicted using dashed lines and the drag atoms are higlighted by empty circles.*

However, since the original view from [BJG*15] was intended for the analysis of molecular tunnel behavior in molecular dynamics simulation, we modified the view in several ways to fit our needs. First, we enable the users to switch between the contour representation of molecular tunnel and the disc representation that is used in the CaverDock algorithm. Second, we display the ligand positions within the Ligand View. For this we offer two alternative views. By default we use orthogonal projection, where we project the atoms and bonds of the ligands to the plane given by the corresponding tunnel disc. For smaller ligands this representation is preferable as it naturally preserves the mutual ligand orientation as well as the positioning of the ligand atoms concerning the surrounding amino acids. However, for ligands with a larger number of atoms and bonds, the orthogonal projection might become too cluttered. Therefore, we also offer a standard 2D structure diagram representation of the ligands, generated by the Chemistry Development Kit library [WMA*17] (see Figure 9). This representation gives a better overview of the ligand structure but does not provide any information regarding the changes in the spatial conformation of the ligands. It is also less precise in terms of distance and orientation between the atoms and surrounding amino acids. To partially address the second issue, we rotate the structure diagrams of the ligands to minimize the difference in distances between the ligand atoms and amino acid in the 3D and 2D representations, as well as the distances between the ligands themselves. In both representations, the hydrogen bonds between the atoms of the ligands and the amino acids are depicted by dashed lines. Additional information about the exact atoms forming the bond is shown on mouse hover. Users can also optionally highlight the drag atom, i.e., the atom that is placed within the tunnel disc. If this option is enabled, the part of the ligand behind the disc is depicted using the dashed lines (see Figure 5). This helps users to determine the actual spatial orientation of the ligand.

4.5. 3D View

The above-mentioned views provide a fast and intuitive way to communicate the most important information about the binding free energy and amino acids in the close vicinity of the ligand during the transportation. However, in order not to overwhelm the user, these views are abstracting from the detailed spatial information. In order not to lose this information, we accompanied the abstract 2D views with the 3D View (see Figure 2e), depicting the protein structure and both lower- and upper-bound trajectories of the ligand in the currently selected frame. The main strength of this view is coming from the fact that it is interactively linked with all other views. This means that the users can always obtain additional spatial information (e.g., position and orientation of amino acids) through interaction.

The 3D view supports all common molecular representations [KKF*17] for both protein and ligand structures. One of the most popular ones is the cartoon representation which depicts the secondary structures and hence it directly enables to study their shape and orientation (**R5**). We decided to communicate this information solely via the 3D view as it is the most straightforward communication channel for this purpose.

The main goal of CaverDock is to investigate ligand transportation from the outer environment to the inner part of the protein structure. Therefore, we need to make sure that the important parts of the data (i.e., ligand and amino acids in its vicinity) are not occluded by the rest of the protein when the ligand gets deep into the protein structure. To handle this issue, we provide the users with an option to turn on a clip plane which removes the unimportant parts of the protein and allows them to "look inside". By default, the orientation and position of this clip plane are set such that they correspond to the disc that was used during the ligand fitting by CaverDock in the currently selected frame. Note that it is the same plane as the one used for computing the projection in the Ligand View. The users can, however, manually modify both position and rotation of the clip plane in 3D as they see the fit.

Finally, to provide the users with the additional context, we also enable to visualize the tunnel that was used by ligand during the transportation (**R4**). Our solution supports both original spherical representation used by CaverDock as well as more realistic surface representation [JBSK15] that can capture the asymmetric shape of the tunnel. In both cases, the users can decide whether the tunnel representation should or should not be affected by the clip plane.

5. Evaluation

To verify the usability and contribution of our tool, it was tested by the senior protein engineer from our collaborating research group. For that we used multiple datasets produced by the CaverDock tool. In this section, we demonstrate the usefulness on a case study, conducted using one of these datasets and describe the most notable feedback we received from the domain expert.

5.1. Case Study

This dataset simulates the transportation of the Adenosine Triphosphate (ligand, ATP) to the active site of Carbamoyl Phosphate Synthetase (protein with PDB ID 1A9X). The ATP molecule has an elongated shape with Adenine (aromatic rings) on one side of the ligand and the phosphate groups on the other side (see Figure 6). To ensure that the ligand reaches the active site, the simulation was performed in the reverse order, i.e., the ligand was placed into the active site and CaverDock was used to compute its movement towards the outer environment. The simulation was performed for three different tunnels present in the 1A9X protein. However, finding the exit trajectory was fully successful only for one of these tunnels.

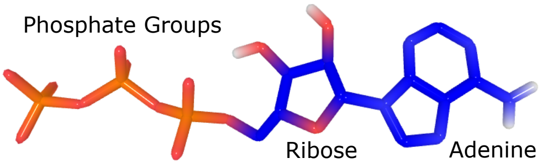


Figure 6: *Adenosine triphosphate (ATP) molecule consisting of Adenine with aromatic rings, Ribose, and the Phosphates groups.*

After loading the data into our tool, the expert started the analysis by looking at the Energy Profiles (see Figure 7). The first thing he noticed was that the RMSD plot contained two areas with large structural differences – around frame 20 and between frames 50 and 80. He also noticed that the changes in RMSD correspond to frames with large conformation changes indicated in the lower-bound trajectory. Therefore, he turned to the 3D View to try to identify what happened in these frames.

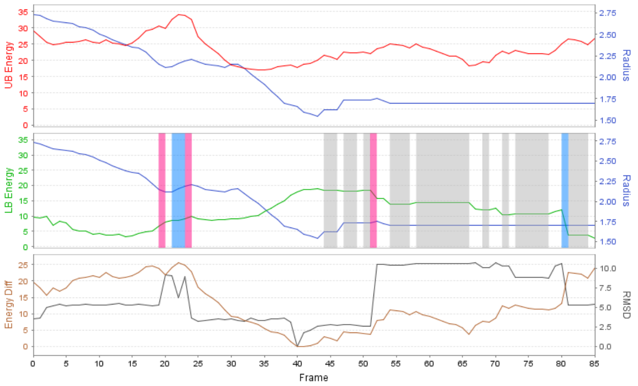


Figure 7: *RMSD (bottom chart, black line) depicts two parts with significant structural differences between the lower- and upper-bound trajectories around frame 20 and between frames 50 and 80. The color stripes in the lower-bound trajectory (middle) show whether these differences are caused by rotation only (blue) or are also accompanied by other conformation changes (pink).*

The expert noticed that at the beginning of the simulation, both ligands were positioned with the Adenine side closer to the protein active site. However, the orientation of the Phosphate groups was different. Upon displaying the main tunnel used for computation of the trajectories, it was apparent that while the ligand from the upper-bound trajectory follows the tunnel, the Phosphate groups of the lower-bound ligand were sticking outside of the tunnel (see Figure 8). The expert suspected the presence of another tunnel at this location. Therefore, he loaded two other tunnels into our tool, which confirmed this assumption. Moreover, from the shape of the tunnel the expert immediately saw that the ATP would not be able to pass through it due to the sharp turn in the middle (see Figure 8).

The expert then decided to explore what was happening in the frames with high conformation changes. He forwarded the trajectories to frame 18, before the first change. Then, by slowly browsing through individual frames in 3D he quickly noticed that the lower-bound ligand completely changed its orientation several times—in some of the frames, the Phosphate groups were oriented towards the active site, while in others it was Adenine. After these changes, the two ligands seemed to align better and followed a similar path until frame 50, were the lower-bound ligand flipped its orientation again and continued moving in this orientation almost until the end of the simulation.

While it was obvious that the drastic change of orientation exhibited in the lower-bound simulation is infeasible for the ligand within the free space of the tunnel, the expert was still curious why the reverse orientation of the ligand was considered more energetically suitable in the simulation. Therefore, he turned to the Lining Residues Plot and the Ligand View to explore the presence of the hydrogen bonds (see Figure 9). He used the Lining Residues Plot to navigate to the frames where the bonds were present, particularly in the parts of simulation where the bonds were detected in multiple consecutive frames. Using the Ligand View and the 3D View, he then studied the precise location of the bonds. In the Ligand View he alternately used the orthogonal projection of the ligand to better estimate the orientation of the atoms, and structure diagram to better see where within the ligand structure the bond was formed. He noticed that several hydrogen bonds were also formed with the lower-bound ligand in the reversed orientation.

Based on his observation from our tool, the expert made two hypotheses. Regarding the part of simulation where one of the ligands seemed to travel through the tunnel that was not intended for this simulation, the expert noted that this was probably due to wrong initial settings of the simulation computation. Normally, the atom at the ligand's center of mass is selected as the drag atom (i.e., the atom that is always placed within the tunnel disc during the simulation), but the expert speculated that for elongated ligands, such as ATP, this might not be the best choice.

The second hypothesis is related to the reversed orientation of the ligand. Based on the large portion of the simulation the ligand spent in this position as well as the presence of the hydrogen bonds which suggests the correct orientation, the expert speculated that it is possible that in the native state the ligand might be oriented such that the Phosphate groups should face towards the active site and that also the initial placement of the ligands in this simulation might have been incorrect.

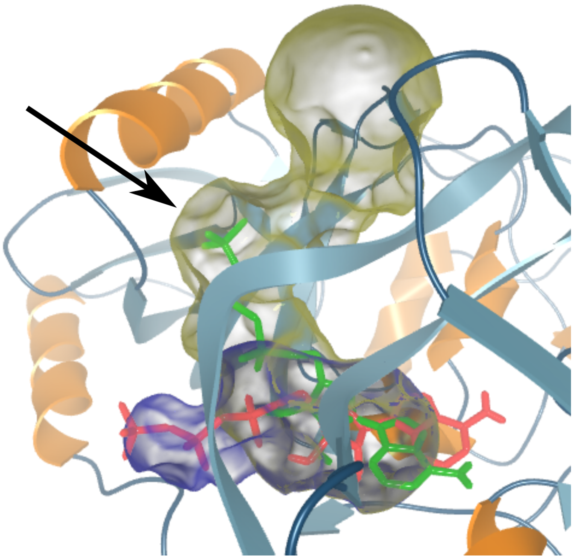


Figure 8: *The ligand from the upper-bound trajectory (red) is following the main tunnel (blue) while the Phosphate groups of the lower-bound ligand (green) are sticking outside of the main tunnel and following a secondary tunnel (yellow). The sharp turn in the middle of the secondary tunnel is marked by arrow.*

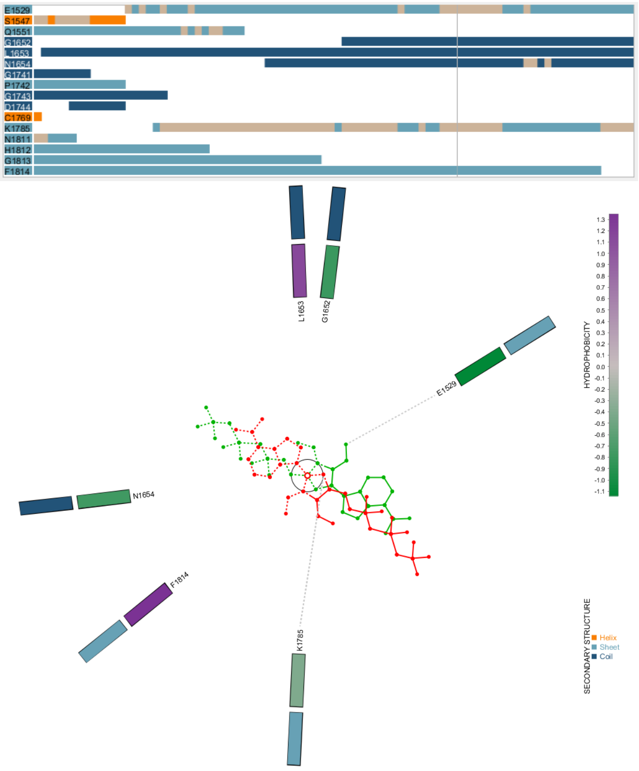


Figure 9: *Lining Residues Plot (top) and Ligand View (bottom) showing the hydrogen bonds (depicted by dashed lines) at frame 60. The ligand is depicted using the structure diagram based on the Chemistry Development Kit [WMA*17].*

5.2. Discussion and Feedback

Here we discuss the feedback we received from the domain expert in the form of informal interviews during and after the testing.

First of all, the expert appreciated the initial alignment of the trajectories with regards to the ligand position in the protein tunnel. This alignment is performed in the data preparation phase. He noted that without our tool he had to manually search for the corresponding frames whenever he wanted to compare the trajectories and their properties at a specific position in the protein tunnel. Thanks to the alignment step, the expert could easily navigate himself through the trajectories and compare them at the specific position.

The expert used the Energy Profiles and Lining Residues Plot as a form of navigation to interesting frames. Here he particularly liked the plot showing the RMSD between the two ligand conformations. He stated that this information was previously inaccessible to him. When focusing on a single trajectory, the expert appreciated the highlighting of frames that were subject to large conformation changes. He also stated that the distinction between simple rotation and conformational changes is very important for his analysis and he was unable to retrieve such information before without tedious manual scripting.

The expert also mentioned that having the direct visualization of hydrogen bonds in both Lining Residues Plot and Ligand View was very helpful in this case as they were instrumental in identifying more probable conformations. He stated that it is highly probable that he would miss them when using PyMol because he would have to know a priori where exactly they are formed to set the visualization properly. On the other hand, he suggested that it would be beneficial to show in the Lining Residue Plot which of the ligands formed the bond at the particular position, as well as the bond count. This information would be beneficial as the higher number of bonds indicates stronger reaction affinity. He also stated that the detailed list of the bonds formed at the selected frame could be shown in the tooltip of this plot. This information can currently be accessed in the Ligand View. However, the user first needs to navigate to a single frame, which makes the exploration more tedious.

In general, the expert clearly stated that he would never be able to make the observations described in the previous section without our tool. He did not look at the simulation in 3D until now as it was too complicated for him to set up the alignment of the trajectories manually using some of the traditional 3D tools.

6. Conclusion and Future Work

In this paper we have presented our proposed system for the interactive visual analysis of molecular docking results. The system enables the domain experts to quickly verify their data and explore its various properties. Although our system operates primarily with data from the CaverDock tool, we believe it can be easily extended for other tools as our system processes a sequence of frames describing positions of atoms, which is a common output of many related tools.

In the near future, we plan to address the remarks collected from the domain expert during the testing. We would also like to focus on better representation of the secondary structures surrounding the tunnel. The current representation is beneficial for the estimation of their role when exploring individual frames. However, it is not sufficient to classify tunnels based on the secondary structures and explore the relationship between tunnel and ligand classes.

7. Acknowledgement

We would like to thank Ondřej Vávra from Loschmidt Laboratories for providing the datasets and valuable feedback. The presented work has been supported by the Czech Science Foundation (GAČR) under research project no. 17-07690S.

References

[ABM08] AMARO R. E., BARON R., MCCAMMON J. A.: An improved relaxed complex scheme for receptor flexibility in computer-aided drug design. *Journal of Computer-Aided Molecular Design 22*, 9 (2008), 693–705. 2

[ABSC12] AL-BLUWI I., SIMÉON T., CORTÉS J.: Motion planning algorithms for molecular simulations: A survey. *Computer Science Review 6*, 4 (2012), 125–143. 2

[BFR00] BISSANTZ C., FOLKERS G., ROGNAN D.: Protein-based virtual screening of chemical databases. 1. evaluation of different docking/scoring combinations. *Journal of Medicinal Chemistry 43*, 25 (2000), 4759–4767. 2

[BJG*15] BYŠKA J., JURČÍK A., GRÖLLER M. E., VIOLA I., KOZLÍKOVÁ B.: MoleCollar and Tunnel Heat Map visualizations for conveying spatio-temporo-chemical properties across and along protein voids. *Computer Graphics Forum 34*, 3 (2015), 1–10. 6

[CLIS10] CORTÉS J., LE D. T., IEHL R., SIMÉON T.: Simulating ligand-induced conformational changes in proteins using a mechanical disassembly method. *Physical Chemistry Chemical Physics 12*, 29 (2010), 8268–8276. 2

[CPB*12] CHOVANCOVÁ E., PAVELKA A., BENEŠ P., STRNAD O., BREZOVSKÝ J., B. K., GORA A., ŠUSTR V., KLVAŇA M., MEDEK P., BIEDERMANNOVÁ L., SOCHOR J., DAMBORSKÝ J.: CAVER 3.0: A tool for the analysis of transport pathways in dynamic protein structures. *PLoS Computational Biology 8*, 10 (2012). 1

[DHR*19] DURAN D., HERMOSILLA P., ROPINSKI T., KOZLÍKOVÁ B., VINACUA A., VÁZQUEZ P.: Visualization of large molecular trajectories. *IEEE Transactions on Visualization and Computer Graphics 25*, 1 (2019), 987–996. 2

[EMA*97] ELDRIDGE M. D., MURRAY C. W., AUTON T. R., PAOLINI G. V., MEE R. P.: Empirical scoring functions: I. the development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes. *Journal of Computer-Aided Molecular Design 11*, 5 (1997), 425–445. 2

[FBM*04] FRIESNER R. A., BANKS J. L., MURPHY R. B., HALGREN T. A., KLICIC J. J., MAINZ D. T., REPASKY M. P., KNOLL E. H., SHELLEY M., PERRY J. K., ET AL.: Glide: a new approach for rapid, accurate docking and scoring. 1. method and assessment of docking accuracy. *Journal of Medicinal Chemistry 47*, 7 (2004), 1739–1749. 2

[FJB*17] FURMANOVÁ K., JAREŠOVÁ M., BYŠKA J., JURČÍK A., PARULEK J., HAUSER H., KOZLÍKOVÁ B.: Interactive exploration of ligand transportation through protein tunnels. *BMC Bioinformatics 18*, 2 (2017). 2

[FVP*19] FILIPOVIČ J., VÁVRA O., PLHÁK J., BEDNÁŘ D., MARQUES S. M., BREZOVSKÝ J., MATYSKA L., DAMBORSKÝ J.: Caverdock: A novel method for the fast analysis of ligand transport. *To appear in IEEE/ACM Transactions on Computational Biology and Bioinformatics* (2019). 1, 2, 3

[Ham02] HAMMES G. G.: Multiple conformational changes in enzyme catalysis. *Biochemistry 41*, 26 (2002), 8221–8228. 2

[HDS96] HUMPHREY W., DALKE A., SCHULTEN K.: VMD: Visual molecular dynamics. *Journal of Molecular Graphics 14*, 1 (1996), 33 – 38. 2

[HEG*17] HERMOSILLA P., ESTRADA J., GUALLAR V., ROPINSKI T., VINACUA A., VÁZQUEZ P.: Physics-based visual characterization of molecular interaction forces. *IEEE Transactions on Visualization and Computer Graphics 23*, 1 (2017), 731–740. 3

[HMOG07] HUEY R., MORRIS G. M., OLSON A. J., GOODSELL D. S.: A semiempirical free energy force field with charge-based desolvation. *Journal of Computational Chemistry 28*, 6 (2007), 1145–1152. 2

[JBSK15] JURČÍK A., BYŠKA J., SOCHOR J., KOZLÍKOVÁ B.: Visibility-based approach to surface detection of tunnels in proteins. In *Proceedings of the 31st Spring Conference on Computer Graphics* (2015), ACM, pp. 65–72. 7

[JFB*19] JURČÍK A., FURMANOVÁ K., BYŠKA J., VONÁSEK V., VÁVRA O., ULBRICH P., HAUSER H., KOZLÍKOVÁ B.: Visual analysis of ligand trajectories in molecular dynamics. In *IEEE Pacific Visualization Symposium 2019* (Thailand, 2019), IEEE. 2

[KKF*17] KOZLÍKOVÁ B., KRONE M., FALK M., LINDOW N., BAADEN M., BAUM D., VIOLA I., PARULEK J., HEGE H.-C.: Visualization of biomolecular structures: State of the art revisited. In *Computer Graphics Forum* (2017), vol. 36, Wiley Online Library, pp. 178–204. 7

[KKL*16] KRONE M., KOZLÍKOVÁ B., LINDOW N., BAADEN M., BAUM D., PARULEK J., HEGE H.-C., VIOLA I.: Visual analysis of biomolecular cavities: State of the art. *Computer Graphics Forum 35*, 3 (2016), 527–551. 1

[KKSE*92] KATCHALSKI-KATZIR E., SHARIV I., EISENSTEIN M., FRIESEM A. A., AFLALO C., VAKSER I. A.: Molecular surface recognition: determination of geometric fit between proteins and their ligands by correlation techniques. *Proceedings of the National Academy of Sciences 89*, 6 (1992), 2195–2199. 2

[KSE09] KORB O., STUTZLE T., EXNER T. E.: Empirical scoring functions for advanced protein- ligand docking with plants. *Journal of Chemical Information and Modeling 49*, 1 (2009), 84–96. 2

[LS11] LASKOWSKI R. A., SWINDELLS M. B.: LigPlot+: multiple ligand-protein interaction diagrams for drug discovery. *Journal of Chemical Information and Modeling 51*, 10 (2011), 2778–2786. 3

[LW99] LIU M., WANG S.: MCDOCK: a Monte Carlo simulation approach to the molecular docking problem. *Journal of Computer-Aided Molecular Design 13*, 5 (1999), 435–451. 2

[MB06] MEILER J., BAKER D.: Rosettaligand: Protein–small molecule docking with full side-chain flexibility. *Proteins: Structure, Function, and Bioinformatics 65*, 3 (2006), 538–548. 2

[Mez03] MEZEI M.: A new method for mapping macromolecular topography. *Journal of Molecular Graphics and Modelling 21*, 5 (2003), 463–472. 2

[MGH*98] MORRIS G. M., GOODSELL D. S., HALLIDAY R. S., HUEY R., HART W. E., BELEW R. K., OLSON A. J.: Automated docking using a lamarckian genetic algorithm and an empirical binding free energy function. *Journal of Computational Chemistry 19*, 14 (1998), 1639–1662. 2

[MLW08] MORRIS G. M., LIM-WILBY M.: Molecular docking. *Methods in Molecular Biology 443* (2008), 365–382. 1

[OSS*14] OHUE M., SHIMODA T., SUZUKI S., MATSUZAKI Y., ISHIDA T., AKIYAMA Y.: Megadock 4.0: an ultra–high-performance protein–protein docking software for heterogeneous supercomputers. *Bioinformatics 30*, 22 (2014), 3281–3283. 2

[PST17] PAGADALA N. S., SYED K., TUSZYNSKI J.: Software for molecular docking: a review. *Biophysical Reviews 9*, 2 (2017), 91–102. 2

[SB98] SHINDYALOV I. N., BOURNE P. E.: Protein structure alignment by incremental combinatorial extension (ce) of the optimal path. *Protein Engineering 11*, 9 (1998), 739–747. 5

[Sch15] SCHRÖDINGER, LLC: The PyMOL molecular graphics system, version 1.8. November 2015. 2

[SLVM08] SAUTON N., LAGORCE D., VILLOUTREIX B. O., MITEVA M. A.: Ms-dock: accurate multiple conformation generator and rigid docking protocol for multi-step virtual ligand screening. *BMC Bioinformatics 9*, 1 (2008), 184. 2

[SM18] SALMASO V., MORO S.: Bridging molecular docking to molecular dynamics in exploring ligand-protein recognition process: An overview. *Frontiers in Pharmacology 9* (2018). 2

[SSVB*13] SEHNAL D., SVOBODOVÁ VAŘEKOVÁ R., BERKA K., PRAVDA L., NAVRÁTILOVÁ V., BANÁŠ P., IONESCU C. M., OTYEPKA M., KOČA J.: MOLE 2.0: advanced approach for analysis of biomacromolecular channels. *Journal of Cheminformatics 5*, 1 (Aug 2013), 39. 1

[TO10] TROTT O., OLSON A. J.: Autodock vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *Journal of Computational Chemistry 31*, 2 (2010), 455–461. 2, 3

[VFP*19] VÁVRA O., FILIPOVIČ J., PLHÁK J., BEDNÁŘ D., MARQUES S. M., BREZOVSKÝ J., ŠTOURAČ J., MATYSKA L., DAMBORSKÝ J.: CaverDock: A molecular docking-based tool to analyse ligand transport through protein tunnels and channels. *Bioinformatics* (2019). 2, 3

[VHG*18] VÁZQUEZ P.-P., HERMOSILLA P., GUALLAR V., ESTRADA J., VINACUA A.: Visual analysis of protein-ligand interactions. *Computer Graphics Forum 37*, 3 (2018), 391–402. 3

[WMA*17] WILLIGHAGEN E. L., MAYFIELD J. W., ALVARSSON J., BERG A., CARLSSON L., JELIAZKOVA N., KUHN S., PLUSKAL T., ROJAS-CHERTÓ M., SPJUTH O., ET AL.: The Chemistry Development Kit (CDK) v2. 0: atom typing, depiction, molecular formulas, and substructure searching. *Journal of Cheminformatics 9*, 1 (2017), 33. 6, 8