Visual Analysis of Ligand Trajectories in Molecular Dynamics

Adam Jurčík
Masaryk University

Katarína Furmanová
Masaryk University

Jan Byška
University of Bergen
Masaryk University

Vojtěch Vonásek
Czech Technical University

Ondřej Vávra
Masaryk University
FNUSA-ICRC

Pavol Ulbrich
Masaryk University

Helwig Hauser
University of Bergen

Barbora Kozlíková*
Masaryk University

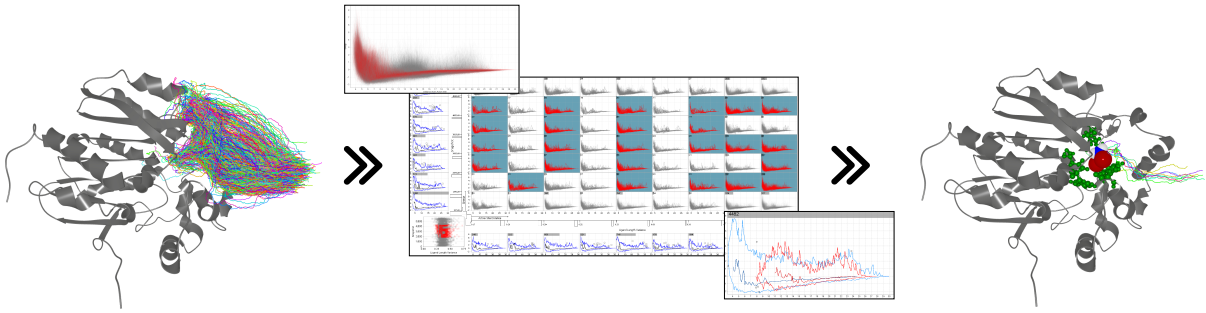

Figure 1: The input dataset, consisting of thousands of possible ligand trajectories, can be analyzed using a set of linked views. In our drill-down approach, the subsets of the original trajectories can be compared and analyzed in more detail, using the combination of chart matrix, selection chart panel, and 3D view.

Abstract

In many cases, protein reactions with other small molecules (ligands) occur in a deeply buried active site. When studying these types of reactions, it is crucial for biochemists to examine trajectories of ligand motion. These trajectories are predicted with in-silico methods that produce large ensembles of possible trajectories. In this paper, we propose a novel approach to the interactive visual exploration and analysis of large sets of ligand trajectories, enabling the domain experts to understand protein function based on the trajectory properties. The proposed solution is composed of multiple linked 2D and 3D views, enabling the interactive exploration and filtering of trajectories in an informed way. In the workflow, we focus on the practical aspects of the interactive visual analysis specific to ligand trajectories. We adapt the small multiples principle to resolve an overly large number of trajectories into smaller chunks that are easier to analyze. We describe how drill-down techniques can be used to create and store selections of the trajectories with desired properties, enabling the comparison of multiple datasets. In appropriately designed 2D and 3D views, biochemists can either observe individual trajectories or choose to aggregate the information into a functional boxplot or density visualization. Our solution is based on a tight collaboration with the domain experts, aiming to address their needs as much as possible. The usefulness of our novel approach is demonstrated on two case studies, conducted by the collaborating protein engineers.

Index Terms: Human-centered computing—Visualization—Visualization application domains—Scientific visualization; Human-centered computing—Visualization—Visualization systems and toolkits;

*e-mail: kozlikova@fi.muni.cz

1 Introduction

Proteins are the essential part of complex mechanisms happening inside the living cells and organisms. They provide a vast variety of functions, such as catalysis of chemical reactions or transport of molecules. In protein engineering, researchers are developing new useful protein variants, based on their understanding of protein design principles. An example can be changing the stability of proteins under normal conditions, outside the lab environment [18], or increasing the protein activity towards other molecules [26].

Proteins are interacting with other molecules (i.e., substrates or ligands) during almost all metabolic processes in cells. These reactions happen in the active site, which is often deeply buried within the protein. The active site has to be accessed by the ligand from protein's exterior. Therefore, the activity and selectivity of proteins with buried active sites is directly influenced by properties of access paths to them [29]. Thus, access paths to the active sites are an important subject to study in protein engineering.

Apart from lab experiments, biochemists employ *in-silico* methods to study protein reactions or transport of ligands. The study of ligand trajectories brings necessary knowledge about important or problematic parts in the paths, which can be possible targets for further optimization by mutagenesis.

When biochemists analyze possible ligand trajectories in MD simulations, they focus on various properties of the ligand, protein, and trajectories, among which they stress the binding energy of the ligand as the most important one. To study these properties and mechanisms in sufficient and reliable detail, they need to run many simulations in parallel. This creates problems not only with storing the data but it also complicates the analysis. Especially when multiple cases of different protein variants or ligand molecules are studied and compared.

It is infeasible to explore datasets in the magnitude of thousands of trajectories using traditionally used molecular visualization techniques. Therefore, the domain experts are forced to design specific methods for the analysis of trajectories individually for each case [16]. Moreover, the parameters for such analysis are usually identified by inefficient trial-and-error approach.

In this paper, we propose a set of interactively linked visualiza-

tions that help to overcome these difficulties. We present a novel system for visual analysis of large ensembles of possible ligand trajectories, whose design was driven by the actual needs of the protein engineers (see Figure 1). They were tightly collaborating on the design of the system, in order to address their needs as much as possible. The proposed system consists of several linked views, enabling the user to explore crucial properties of thousands of ligand trajectories that can be generated by different state-of-the-art simulation algorithms [6, 10, 12, 25].

In the following sections, we first discuss the initial requirements that a visualization tool for exploration and analysis of ligand trajectories must support. Then we describe techniques and methods related to our approach. This is followed by the description of the design decisions and the tool itself. In the end, we demonstrate the usefulness of our tool on two case studies that were performed by our collaborators from the protein engineering laboratory.

2 Data and Task Abstractions

When the biochemists are studying reactions between proteins and ligands, they are usually trying to answer the following questions:

- Q1 Did the ligand get all the way to the active site?
- Q2 Are there any geometrical or energetic obstacles on the ligand's path?
- Q3 What kind of movements of the protein help the ligand to get through to the active site and how do they impact the transport?
- Q4 Does the ligand need to change its shape in order to get to the active site? If so, how?
- Q5 What is the trend in the binding energy of the ligand throughout the whole simulation?

To answer these questions, the biochemists are often using in-silico methods to simulate possible trajectories of the ligand, leading from the outer environment to the protein active site. In general, a ligand trajectory can be described as a set of consequent ligand configurations (i.e., positions and rotations). This data can be produced in various ways, spanning from molecular dynamics (MD) simulations [12, 25], over molecular docking [10], to robotics-inspired algorithms [6].

Due to the stochastic nature of these computational methods, it is necessary to produce large ensembles of thousands of trajectories to draw statistically significant conclusions. As the data is too large to be explored manually by the user, an automated analysis method has to be employed. This often means designing a case-specific measure to cluster the data [16]. To illustrate the approach, we describe several measures that the biochemists usually consider:

- **Ligand Binding Energy** represents the stability and probability of a given ligand-protein configuration. The high binding energy indicates that the ligand is unlikely to bind to the protein in a given location. It can also be indicative of a barrier – physical or chemical – preventing the ligand from passing through the protein void towards the active site. Therefore, there is a strong incentive to focus on conveying the binding energy when exploring the trajectories datasets.
- **Area Under Curve** is defined as the integral of the energy profile along the trajectory. It provides the user with the information about the overall binding energy of ligand through the whole trajectory. High values can indicate that this trajectory is not feasible.
- **Distance to the Active Site** is usually measured as the Euclidean distance between the active site and current ligand position. In many cases, the ligand gets stuck in some part of the protein and it is unable to continue towards the active site. Splitting the data based on this property can reveal the trajectories in which this behaviour occurred.
- **Energy Elevation** – unlike the *Area Under Curve*, which represents the sum of energy along the whole trajectory, the *Energy Elevation* represents the cumulative energetic elevation gain of the ligand. In other words, it shows the sum of energetic ascends along the given ligand trajectory. It describes the overall energetic barrier the ligand has to overcome on its way towards the active site.
- **Structure RMSD** (Root Mean Square Deviation) represents the spatial difference of the protein structure to the mean conformation of this protein. This property helps to identify parts of the data where the protein significantly changed its structure, compared to the average conformation. Such changes can lead to opening or closing of the entrance path to the active site and can thus significantly influence protein reactivity with other molecules.
- **Ligand Conformation Length** gives the biochemists the information about how much the ligand changed its shape along the trajectory. These changes are tightly related to the bottlenecks the ligand encounters on its way towards the active site.

Analysis of these measures of trajectories in a traditional way (i.e., plotting individual scatterplots) often leads to a trial-and-error process with multiple unsuccessful attempts, since the analysis method is set up with only a limited knowledge about the data in question. Usually only simple characteristics of the ligand-protein complex (e.g., RMSD) are evaluated in advance. Worse, the simulation data captures a complex process where multiple events usually occur (e.g., ligand blocking, ligand/protein conformation change, etc.). Therefore, even if the data exhibits high similarities among the trajectories, it cannot be expected that one measure exists such that it describes the whole dataset sufficiently. Instead, it is common that multiple clusters of similar trajectories emerge in the dataset.

In general, the ligand trajectory data can be understood as a set of functions describing different properties. For example, the ligand positions and rotations along the trajectory can be described by a vector-valued function of time with 6D range. On the other hand, the *Distance to the Active Site* can be described by a function with 1D range. The existing literature [14, 15, 17] in the field of visualization of functional data suggests that the discussed difficulties, such as data complexity, can be overcome by means of interactive visual analysis.

In order to identify the requirements for the visualization techniques, we have conducted several informal interviews with our collaborators from the protein engineering group. Together we identified the following set of crucial tasks that a visualization system for the exploration and analysis of ligand trajectories must support:

- **T1**: Exploration of the spatial properties of a large number of trajectories in the context of their surroundings (Q1, Q2).
- **T2**: Exploration of large sets of functions describing physicochemical properties along the trajectories (Q2, Q4, Q5).
- **T3**: Ability to easily reveal parts of the trajectories exhibiting different behaviour than the rest (Q2-Q5).
- **T4**: Possibility to identify critical properties by dividing trajectories with similar behaviour into clusters (Q1, Q3, Q4).
- **T5**: Capability to identify differences (both spatial and physicochemical) among clusters of similar trajectories and relate these findings to the possible cause in the surrounding environment (Q2, Q3).
- **T6**: Possibility to compare trajectories from multiple datasets (Q1-Q6).

3 Related Work

Visualization of biomolecular structures has been a research topic already for a couple of decades. Much work has been put into the improvement of molecular representations and providing insight into molecular structure and properties at various levels of detail [19]. In terms of protein-ligand interactions, a substantial work [21] is focused on the exploration of protein cavities, which play a crucial role in protein reactivity.

Other tools focus directly on the analysis of the process of protein-ligand docking. From this group, we can mention LigPlot+ [22], which offers a way to compare multiple protein-ligand docking conformations via 2D plot representation. However, this tool does not support dynamic data. Hermosilla et al. [13] tried to address this issue and presented a system for analysis of binding forces in a protein-ligand docking simulation. In this system, the users can interactively explore the position of the ligand and the chemical forces taking place during the molecular docking in annotated 2D and 3D views. However, the tool focuses on the exploration of single protein-ligand docking simulation. Moreover, as it animates the dynamic behavior, it is not suitable for very large simulations where the observation of the animation is not feasible anymore.

Another approach to the analysis of protein-ligand docking is the visual exploration of molecular trajectories in the MD simulations. Furmanová et al. [11] proposed a system which focuses on the analysis of single ligand trajectory, based on the set of physico-chemical properties extracted from the MD simulation. Another tool, VIA-MD [31], offers the users a way to define and extract their own properties from the MD simulation, which can be further interactively explored. It employs density volume in 3D view to indicate interesting parts of the simulation, based on the user-defined settings. However, it is not suitable for simultaneous exploration of individual non-spatial properties of large number of trajectories. Recently, Duran et al. [8] presented another system for exploration of large molecular trajectories. This system is mainly focused on the analysis of binding energies during the MD simulation for which the authors utilize 2D energy plots. These can be compared for multiple trajectories, however, the system was not designed (and thus does not scale) to a high number of trajectories, as each trajectory has its own 2D plot. Another similar tool was proposed by Vázquez et al. [39]. Here, the authors utilize an abstracted interactive representation of trajectory properties accompanied by simulation energy plot. Similarly, the tool supports comparison of several trajectories but it is efficient only for a very limited amount of timesteps.

Another example of analysis of trajectory datasets in the molecular context can be found in several systems, focusing on the exploration of flow of water molecules in MD simulations. These usually represent substantially larger datasets, consisting of thousands of trajectory ensembles. Bidmon et al. [3] presented a clustering-based method for extraction of principal paths, which are then represented as tubes in the 3D context. Vad et al. [35] employed density isosurfaces to represent the protein inner space occupied by water molecules, as well as a set of abstracted 2D representations for more thorough exploration and filtering. Vassiliev et al. [38] proposed a method for extraction and visualization of water diffusion within protein, based on a streamline tracing algorithm used in fluid dynamics. AQUA-DUCT [23] employs clustering based on protein entry points of water molecules and then visualizes the resulting paths. Another example of tracking large molecular flow data can be found in MolPathFinder [1]. Here the authors track and visualize flows of atoms in MD simulations. Similarly, Chavent et al. [5] presented a system for analysis of lipid movements. All of these systems, however, focus on the analysis of general flow trends. Although the similar principles could be partially applied to ligand trajectories, in the current stage they do not answer questions specific to protein-ligand interactions, such as *what is the ligand binding energy at a specific location in the protein* or *if (and how) the ligand changes its shape during the MD simulation*. The same stands for many other trajectory analysis techniques from different domains.

Nevertheless, we can find several trajectory analysis techniques used in other domains, which can be successfully applied in our case. Curve Boxplot [24] is a method of deriving robust and descriptive statistical information from an ensemble of multivariate curves. Vad et al. [36] adapted this technique for root growth ensembles. A similar approach was used by Ferstl et al. [9] for conveying statistical properties of streamlines passing through a selected location from an ensemble of flow fields. Here streamlines are transformed into a low-dimensional Euclidean space, clustered into major trends, and then approximated by a multivariate Gaussian distribution. A method for spatial generalization and aggregation of trajectories representing movement data was introduced by Andrienko and Andrienko [2]. Their method transforms trajectories into aggregate flows between appropriate areas by extracting significant points. Demšar and Virrantaus [7] discussed the concept of 3D space–time density of trajectories to solve the problem of cluttering in the space–time cube.

From the methods described above, it is evident that in most cases the trajectory data has to be grouped in order to present it in an informative way. This is usually done based on a set of properties, either extracted from the trajectory records (e.g., the curvature of the trajectory, velocity of the moving object), or supplied as additional metadata (e.g., binding energy). Pobitzer et al. [28] examined trajectory attribute space and, using exploratory factor analysis, identified a set of six expressive properties that describe the patterns in large sets of particle path lines. However, while these trajectory features are certainly important, they do not take into account forces and properties that are specific to the molecular domain. These properties play a crucial role in protein-ligand docking and are of utmost importance to the domain experts. It is thus inevitable to facilitate the exploration of these properties and to consider them when grouping the trajectories. Such properties can be represented as functions of the time variable and further explored. Since the same stands for molecular data, we provide an overview of methods that can be used for analysis of large sets of functions.

Jacques and Preda [15] presented a survey of functional data clustering methods, in which they recognize four commonly used methodologies: *raw-data methods* working directly on the evaluation points of the curves, *filtering methods* which first approximate the curves into a finite basis of parameters and second perform clustering, *dimensionality reduction methods*, and *distance-based methods*. However, the clustering algorithms suffer from some well-known problems. Many of them require a priori knowledge of the data, e.g., in order to define the number of clusters or to set up the clustering parameters correctly. Moreover, in case of biochemical data, it is crucial for the user to understand what leads to the formation of certain clusters and how the changes of clustering parameters influence the result. Then a feasible alternative is the user-controlled data binning in combination with other interactive visual analysis techniques. For example, van den Elzen and van Wijk [37] proposed a concept of small multiples (originally introduced by Tufte [34]) and large singles for interactive exploration of multivariate data. Konyha et al. [17] use a similar linked-view approach in a tool for interactive analysis of families of function graphs. Piringer et al. [27] also use multiple coordinated 2D and 3D scatterplots and histograms. In general, we can conclude that the concept of multiple coordinated views is nowadays well established in visualization literature and is understood and used by experts from different application domains.

Many publications are focusing also on visual representation and analysis of protein cavities, i.e., the void space which can be potentially used for transportation of a ligand. A comprehensible overview of these techniques was published by Krone et al. [20]. These methods mostly convey the shape of the cavities, their evolution over time, and the physico-chemical properties of their surroundings. However, they do not deal with the ligand trajectories themselves.

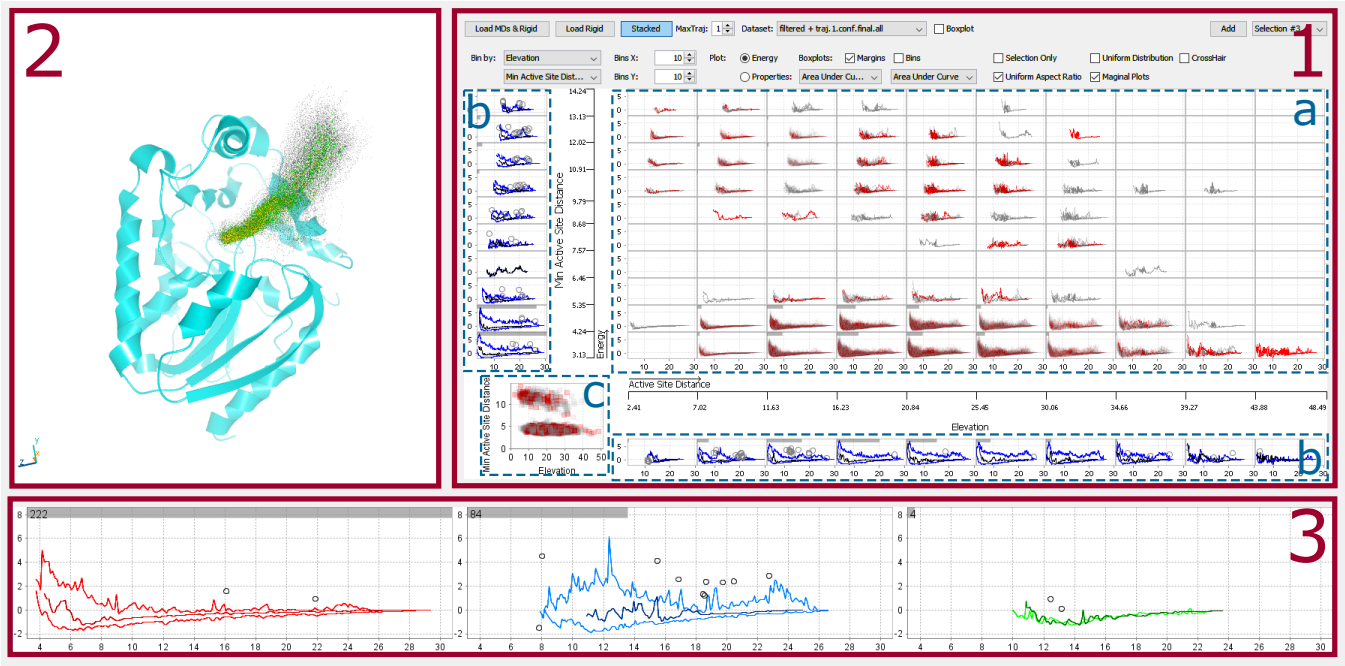

Figure 2: Overview of the proposed tool and its individual parts: Chart Matrix (1), 3D View (2), and Selection Chart Panel (3). The letters mark the individual parts of the Chart Matrix: Small Multiples View (a), horizontal and vertical Marginal Charts (b) and overview scatterplot (c).

4 Visual Analysis of Trajectories

The above described requirements (**T1**-**T6**) guided the design of our proposed visualization system which employs well-established visualization concepts in several interactively linked views (see Figure 2). Together, these views support the exploration and comparison of datasets consisting of thousands of trajectories.

The size of our input trajectory dataset intrinsically requires different levels of visual abstraction for its exploration. At the most abstract level, we represent each trajectory by a set of scalar values that are derived from geometrical or physico-chemical properties, describing the whole trajectory at once (see Section 2). These properties we denote as the *binning properties*. This level of abstraction enables us to divide trajectories into clusters exhibiting similar behaviour with respect to their binning properties (**T4**). The second level of abstraction comes from the need to represent some physico-chemical properties along the ligand trajectory (**T2**) in more detail. We call them *profile properties* and represent them in form or functional data which allows us to use several techniques, such as functional boxplots, to communicate the main trends and changes (**T3**) along the trajectories. In the context of this paper, we employ only the ligand binding energy as it is the most important profile property for our collaborators. However, the concept can be easily generalized for any property that can be measured along the trajectory.

To utilize the profile properties from the second level of abstraction to compare clusters of similar trajectories (**T5**) identified on the first abstraction level, we employ the concept of small multiples [34] in our proposed Chart Matrix (Section 4.1). The Chart Matrix enables the users to drill down through the data in an informed way, further eliminating the number of trajectories and keeping only those that are relevant to the user's task. It is interactively linked with the 3D View (Section 4.2) which shows the spatial correspondence between trajectories and the protein (**T1**). Moreover, the Chart Matrix enables the users to create and store trajectory selections, including selection from multiple datasets, which can be explored and compared (**T6**) using the Selection Chart Panel (Section 4.3).

4.1 Chart Matrix

The proposed Chart Matrix consists of several components: *Small Multiples View*, *Marginal Charts*, and *Summary Scatterplot* (see Figure 2). The *Small Multiples View* (see Figure 2, section 1a) tackles the challenge of exploring and analyzing large sets of ligand trajectories and their profile properties (**T2**) by splitting them into the matrix of bins, according to two user-defined binning properties. The value range of each selected binning property is split into n equal parts mapped to columns and rows of the matrix, where n is defined by the user separately for each of the two binning properties. Each chart C_{ij} in the Small Multiples View then depicts a selected profile property (e.g., the ligand binding energy) for a subset of trajectories that were assigned to bins i and j according to the binning properties.

The smaller chunks of data are easier to analyze while the binning can reveal interesting trends, patterns, and correlations between the two selected binning properties (see Figure 2, section 1a). In order to allow the users to identify the values of the binning properties possessed by the most of the trajectories as well as the data outliers, we equipped each chart with a bar indicating how many trajectories fell into a given bin, i.e., how many energy profiles the chart depicts.

In order to provide the users with easier navigation and ability to identify binning properties that cluster the data (**T4**), the side axes of the main matrix area are equipped with *Marginal Charts* (see Figure 2, section 1b). These contain the summary of trajectories depicted in individual rows and columns. In other words, they split the data based on a single binning property, corresponding to the axis where they are located. The Marginal Charts are also equipped with the bar indicating the size of the contained data.

Finally, the *Summary Scatterplot* in the lower left corner of the Chart Matrix (see Figure 2, section 1c) shows the overview of the whole dataset. The axes of the scatterplot are formed by two currently selected binning properties. Thus the arrangement of the data in the scatterplot reflects the arrangement of the data in the Small Multiples View.

Profile Property Charts

Since the charts in both the Small Multiples View and Marginal Charts view can contain a large number of trajectories, the users can depict the profile properties using the functional boxplot [32]. This method reduces the visual clutter, while providing the information about the distribution which allows them to easily identify trends and areas among the aggregated trajectories exhibiting different behaviour (**T3**).

Alternatively, in cases when it is necessary to examine individual trajectories, but some of the bins still contain large sets of data, we employ the logarithmic opacity modulation based on the number of trajectories depicted in the chart. When the chart contains a large number of energy profiles, these are rendered with low opacity, which is accumulated in areas where the profiles are overlapping. This way the main energy profile trends are visible even in the charts with a large amount of data entries. The data in sparsely populated charts are assigned high opacity values and thus also the individual energy profiles in these charts are clearly visible.

In order to easily compare the profile properties (e.g., energies) among individual clusters of trajectories (**T5**), we ensure that all charts use the same axes ranges. Moreover, we employ a crosshair navigation, which points to the same coordinates in each of the plots. Nevertheless, it may still be hard to compare details among the bins, due to the often very small size of the individual plots. Therefore, we introduced uniform zooming across the charts – when the user zooms into one plot, all the remaining plots, including the Marginal Charts, zoom into the same data interval. This facilitates the exploration and comparison of the close-ups of the profile properties.

Drill-Down

The Chart Matrix also offers several ways of data selection and drill-down approach. It is possible to use brushing in the Small Multiples View or Marginal Charts to select a rectangular region of charts or to individually pick multiple charts (see Figure 1). The data entries from the brushed charts are then added to the selection. Alternatively, the users can also utilize the Summary Scatterplot for selection of trajectories using rectangular brushing. Upon creating a selection, the charts in the Small Multiples View and Marginal Charts are recomputed to depict only the selected subset of the data. This enables the user to interactively and iteratively drill down through the trajectory data and explore the selected trajectories in more detail. At any point in time it is also possible to adjust the number of bins per axis – e.g., when the number of displayed data entries no longer requires splitting into numerous bins – or change the binning property.

As the user can choose to display only specific parts of the data, we use the Summary Scatterplot (which always depicts the whole dataset) to provide clues about the portion of the data that is currently selected and visible (see Figure 2, section 1c). The Summary Scatterplot also enables the users to observe the distribution of the selected data within the scope of the whole dataset. This can be particularly useful in cases when the change of the binning properties (also used for data mapping in the scatterplot) reveals the new patterns in the data.

Note that the user can choose to keep the whole dataset visible at all times also in the Small Multiples View. In such case, the energetic profiles of the selected trajectories are highlighted in the individual plots (see Figure 2, section 1a) and the users can observe the aforementioned distribution changes also in the Small Multiples View.

Alternative Binning

While the default approach based on the uniformly split range intervals enables the users to identify the properties dividing trajectories into clusters (**T4**), it can also lead to over-plotting in some of the data bins, while some other parts of the plot matrix remain empty. The empty or almost empty blocks help to visually cluster the data based on the selected properties which is important at the beginning of the analysis but becomes less important as the users obtain a better understanding of the data. Therefore, we also provide the users with a possibility to split the data into uniformly populated bins. This way we reduce the over-plotting and offer a more screen-space efficient way of displaying the data. To achieve this, we sort the data items by one of the selected properties and then split them into n uniformly populated bins. We then compute the threshold property values between individual bins. In the same way, we also compute threshold values for the second binning property. We then use the computed thresholds to split the data in the Small Multiples View. Naturally, since we combine the two selected properties for binning, the data population in the individual charts will not be completely uniform, but this method does yield nearly uniform results and sufficiently fulfills the requirement for a more screen-space efficient method.

Linked Views

The selection in the Chart Matrix is tightly connected to other views. The selected trajectories are also depicted in the *3D View* (see Section 4.2) which allows the users to explore the subset of similar trajectories in the context of their surroundings (**T5**). The users can also store the selection at any point in time and explore it later in the Selection Chart Panel (see Section 4.3).

Another interaction possibility, also tightly bound to the *3D View*, is the exploration of a single energy profile (**T3**). On mouse hover, individual energy profiles in the Chart Matrix plots are highlighted. Clicking on a point of the highlighted energy profile navigates the camera in the *3D View* to the corresponding part of the trajectory, where the ligand and its surrounding amino acids are highlighted. This enables the users to explore the interesting parts of the energetic profile in the context of the protein structure (**T1**) and thus to better assess, e.g., the causes of high energy peaks in the explored profiles.

4.2 3D View

While the above described Chart Matrix provides a powerful tool for the data exploration, the actual spatial characteristics of the explored trajectories are highly abstracted. As the exploration of the spatial properties is required by the tasks **T1** and **T5**, we remedy this drawback by providing the users with a 3D view (see Figure 2, section 2) where the currently selected trajectories can be visualized in the context of the relevant protein molecule. This view supports all common representations of molecules [19]. However, in the context of the exploration of ligand trajectories, we found out that it is most useful to use non-space filling representations that depict only the molecular architecture, such as the *cartoon* representation [30].

When visualizing the ligand trajectories in 3D, we are facing two problems. First, the number of trajectories that needs to be depicted can be enormous and the naïve approach (i.e., simply depicting all of them) suffers from visual clutter and occlusion problems. The second issue comes from the dynamic nature of the data where the trajectories represent the ligand movement over time. Hence, there is no single reference molecule as the protein is changing its shape over time as well.

We overcome the first challenge by utilizing a density visualization instead of showing all individual trajectories. We estimate the density function based on the number of trajectories passing through a given part of space. In other words, we compute a voxel grid where each cell contains a number of trajectories passing through it. For the purpose of our prototype, we use a simple glyph-based visualization of this grid (see Figure 3). We represent each cell by a sphere where its size and color are modulated based on the user-defined values, corresponding to the percentiles in the estimated density. To suppress artifacts coming from the uniform grid layout we apply a small jitter to the sphere positions. We acknowledge that the state-of-the-art techniques (e.g., streamline variability plots [9])

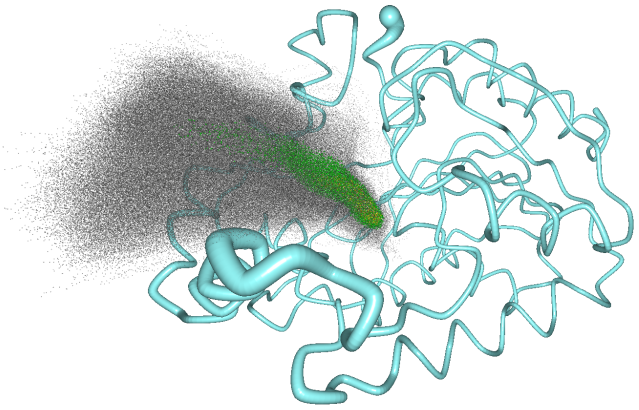

Figure 3: Portion of the LinB dataset (>13,500 trajectories) visualized using our aggregation approaches. Trajectories are depicted using the density based visualization. The structure is shown using the protein backbone with its width modulated by the maximum extent w.r.t. its mean conformation.

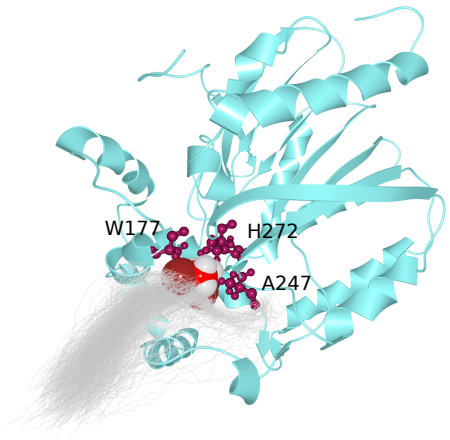

Figure 4: Ligand at its nearest site w.r.t. the active site where it was transported by following a trajectory. Surrounding amino acids that prevented the ligand from further movement towards the active site are depicted using *Balls and Sticks* visualization and labeled.

would produce more visually "pleasing" images, but our aim was to use a simple visualization and prove the concept. In order to fully address the task **T1**, we also provide the users with the standard representation of individual trajectories using 3D curves for cases when the number of selected trajectories that needs to be depicted is low enough (see Figure 4).

To address the second challenge, dealing with multiple conformations of the protein structure over time, we utilize a concept of *mean molecule* that is often used by the domain experts. As the name suggests, the mean molecule is created by averaging all atoms positions over time. The problem is that such representation does not provide any information about the actual protein dynamics. Therefore, we encoded the flexibility of each atom (i.e., we employ the maximal possible difference from the mean in the set of snapshots) to the radius of the tube in the alpha-carbon trace representation (see Figure 3). The backbone representation was chosen because the applied tube modulation is easier to perceive than the same technique applied to the cartoon representation.

The 3D view provides standard means for the user interaction, such as rotation, panning, and zooming. Additionally, in order to better support the task of ligand trajectory exploration (**T1**), we implemented a feature that allows the users to navigate themselves among individual conformations of the ligand along its trajectory, including automated replay.

4.3 Selection Chart Panel

The one requirement which was not addressed yet is the comparison of multiple datasets (**T6**). The Chart Matrix, described above, supports the exploration of only one dataset at a time which makes it impractical for comparison of multiple datasets. The reason why we decided not to display multiple datasets in the Chart Matrix at the same time is twofold. First, the datasets tend to have very different distribution of the data. Second, each dataset can often consist of tens of thousands of trajectories. If we would combine the datasets, both these reasons would make the Chart Matrix view more cluttered, which would lead to more complicated exploration of the data. Instead, we let the users to evaluate each dataset individually so that they can more easily understand the underlying features in each dataset.

Once the users build a mental picture about the individual datasets they can create one or more selections of the trajectories of interest from both datasets in the Chart Matrix (see Section 4.1). These selections can be compared in a separate view, called the Selection Chart Panel (see Figure 2, section 3). The advantage of the separate view is that more screen space can be devoted to each selection as their number is usually limited.

The charts in this panel support the same interaction and features as those in the Chart Matrix, including highlighting of individual trajectories, shared crosshair and zooming, as well as the possibility to visualize the energy profiles using boxplots or line charts. Additionally, the users can choose to depict the selections either side-by-side or superimposed (see Figure 5). The Selection Chart Panel also enables the users to navigate back to already stored selections such that they can be further explored and refined in the other linked views.

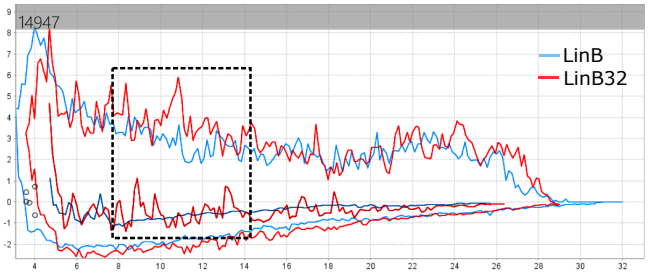

Figure 5: Comparison of trajectories from LinB and LinB32 datasets that transported the DBE to the active site. Noticeable energy peaks can be observed for the LinB32 trajectories at the distance of 8-14 Å from the active site.

5 Evaluation

All the presented techniques were designed in tight cooperation with our collaborators from the protein engineering laboratory. In the end, they were asked to use our tool to explore their datasets and to evaluate its benefits and drawbacks. The evaluation was performed by two post-doc researchers and one doctoral student, all of them focusing on analysis of protein functions and their dynamic behavior.

For the purpose of the evaluation, they employed two datasets consisting of molecular dynamics (MD) simulations of haloalkane dehalogenase LinB (PDB ID 1MJ5) and its mutated variant LinB32 (PDB ID 4WDQ). In order to obtain the ligand binding trajectories,

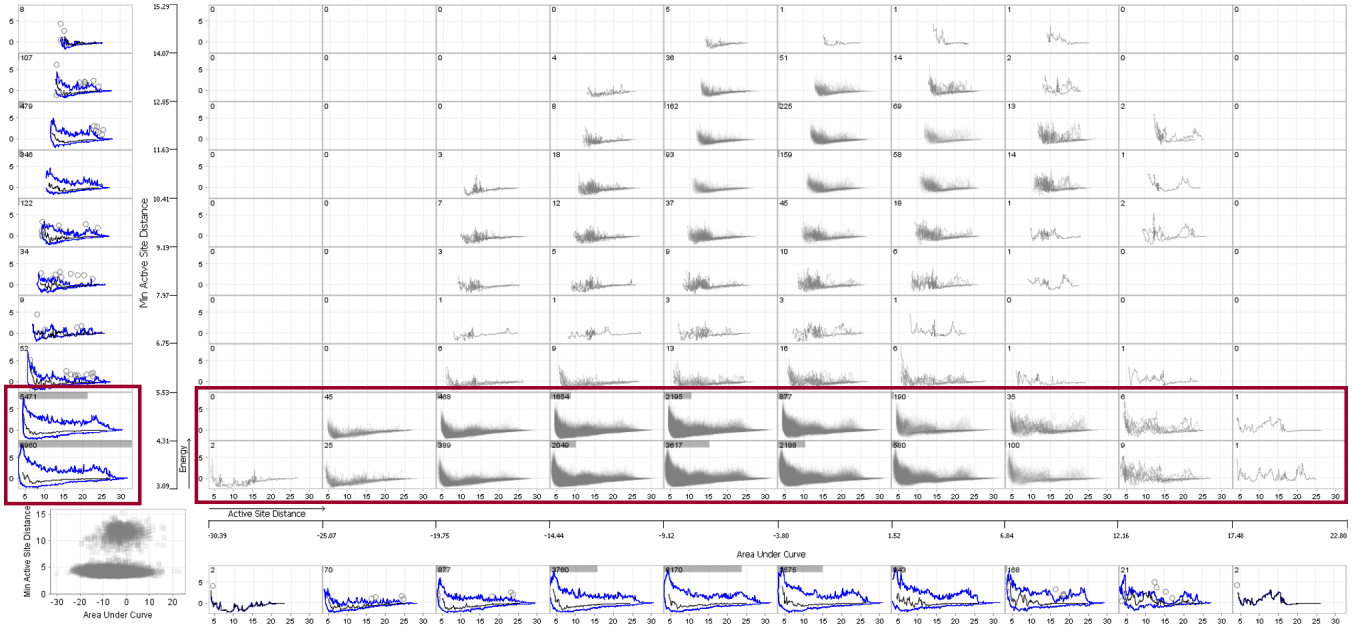

Figure 6: Overview of the LinB dataset visualized using the Chart Matrix. The *Minimum Active Site Distance* and *Area Under Curve* properties were used to slice the data into 10 by 10 bins. The distribution of the trajectory data can be clearly observed. From the bins highlighted by the red rectangle, it can be seen that most of the trajectories transported the ligand at the distance of 3.1-5.5 Å, i.e., to the location near to the active site.

the biochemists computed 15,888 (LinB) and 19,230 (LinB32) trajectories using a path-planning based algorithm. The trajectories were computed on a per-snapshot basis, i.e., the algorithm computed a single trajectory for each snapshot of the protein simulations in an isolated manner. Then the biochemists evaluated these trajectories in terms of energy, using the AutoDock Vina tool [33]. Finally, they enhanced the data with the properties that they usually employ during their standard analysis of ligand trajectories workflow. The following trajectory properties were added: binding energy, area under the curve of the energy, energy elevation, minimum distance to the active site (w.r.t. the whole trajectory), and protein's RMSD.

In the following sections we first present two use case studies and later mention the most important comments of the experts. In the first study, the trajectories of ligand (DBE) binding to two proteins (LinB and LinB32) were explored. This included the comparison of transportation of the DBE in both datasets with respect to the binding energy. In the latter study, the trajectories that transported the DBE to the active site of the LinB were further explored, and a conformation change of the LinB was observed.

5.1 Decreasing Activity of LinB32 (Compared to LinB)

In this study, the main differences among trajectories from the LinB and LinB32 datasets were explored and the biochemists were aiming to reveal the reasons for these changes, using our tool.

First, the domain experts wanted to examine the ratio between the trajectories that reached the active site and those which did not in both input datasets to see if the tool can help with predicting the protein activity. They started with the LinB dataset and tried to retrieve this information using the Chart Matrix view. They used the default 10 by 10 charts layout and chose the *Minimum Distance to Active Site* (rows) and the *Area Under Curve* (columns) properties to divide the data into bins. This setting provided them with an overview of the distribution of the trajectories in terms of energy and their success in reaching the active site. It was possible to see that the vast majority of trajectories from the LinB dataset transported the DBE to the protein active site. This information was easily observed from the small gray bars in the vertical Marginal Charts, sliced according to the *Minimum Distance to Active Site* (see Figure 6). When the domain experts repeated the same process for the LinB32 dataset (see Figure 7), they observed that only a small portion of all the trajectories transported the DBE to the active site. These findings were in agreement with the wet lab experiments showing the binding and catalysis of the DBE in LinB32 to be about 4x slower compared to LinB [4]. This confirms that our tool can be used for predicting protein activity by identifying clusters of trajectories with similar behaviour (**T4**) in faster and less expensive manner than the wet lab experiments.

Naturally, a follow-up question arises about the cause of the lower protein activity observed in the LinB32 dataset. To answer this question using our tool, the biochemists wanted to identify several representatives from all trajectories that could be explored in 3D. They started by looking at large clusters of similar trajectories from which they could choose representatives for further exploration. For this purpose, they again utilized the vertical Marginal Charts, sliced according to the *Minimum Distance to Active Site* (see Figure 7). Using these plots, it was easy to identify a bin consisting of about 25% of all trajectories. These trajectories achieved to transport the DBE only to the distance of 11.7 ± 0.6 Å from the active site. The domain experts further observed that the neighboring bins also contain a significant amount of trajectories. As they were now interested only in the trajectories within these specific bins, they marked them in the Chart Matrix using a rectangular selection tool, and the matrix was recomputed to show only those trajectories. The biochemists decided that they can now decrease the number of bins to 5 rows by 6 columns, since they were exploring only a portion of the whole dataset (see Figure 1 in the Supplementary Material). The energy profiles of individual trajectories in the largest bin were quickly inspected using individual trajectory highlighting on mouse hover. After that, the domain experts selected one representative trajectory and by clicking at it, they requested a 3D visualization of the ligand and its surrounding amino acids at the end of the trajectory. They repeated the process for other bins that contained at least 200 trajectories. In each case, it was possible to observe very similar sets of the surrounding amino acids with the W177 appearing most often

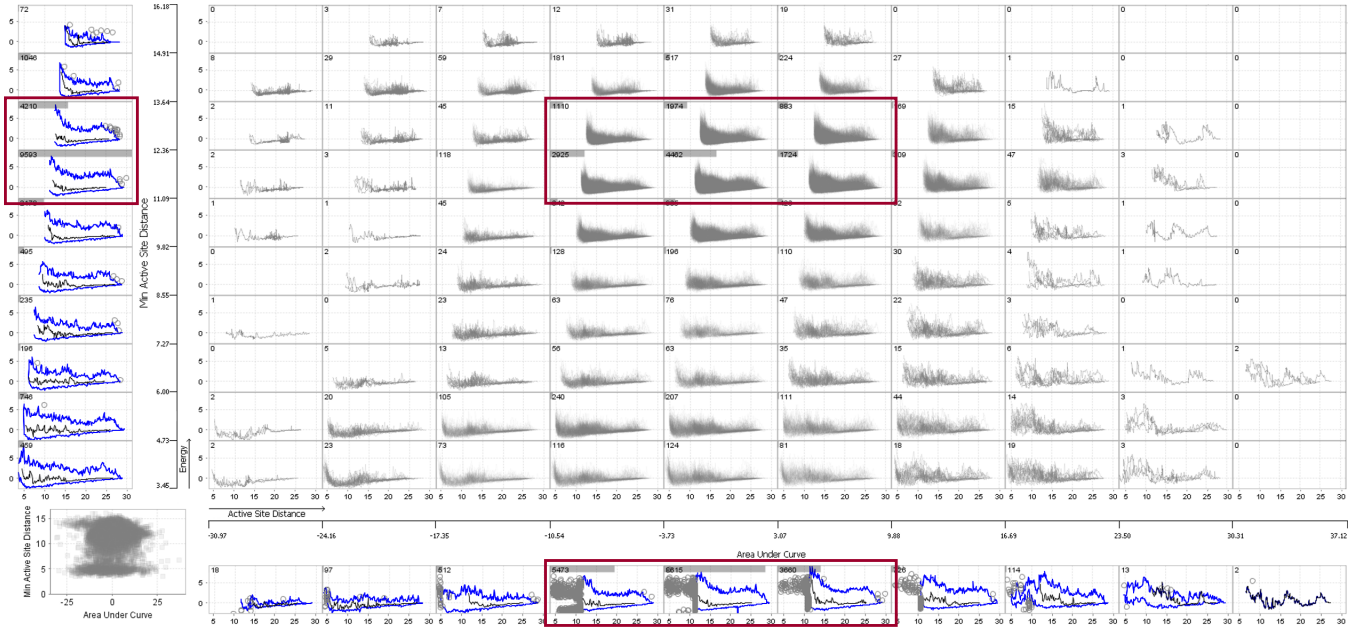

Figure 7: Overview of the LinB32 dataset visualized using the Chart Matrix. The *Minimum Active Site Distance* and *Area Under Curve* properties were used to slice the data into 10 by 10 bins. The distribution of the trajectory data can be clearly observed. From the bins highlighted by the red rectangle, it can be immediately seen that most of the trajectories did not transport the ligand to a location near the active site.

(see Figure 4). From the 3D view, it was obvious that the W177 was blocking the access path.

Also in this case, the results were in agreement with the wet lab experiments as the LinB32 protein was obtained from the LinB protein by mutating the L177 amino acid to W177, in order to decrease the activity of the enzyme by blocking its main access path [4]. This confirmed that our tool enables to explore individual trajectories (**T3**) and to relate the important properties of these trajectories (e.g., the inability to transport the ligand to the active site) to a possible cause in protein structure (**T1**). The domain experts also concluded that our tool would allow them to design such mutation much faster as they want to first validate the possible candidates for mutagenesis in-silico before performing the actual experiments.

Finally, the biochemists posed a question, which they were not able to easily answer before using a single tool: *Is there a difference between the trajectories from both datasets that transported the ligand to the active site in terms of their energy profiles?* To answer this question, they exploited the ability of our tool to easily create selections and compare them. The biochemists started by visualizing the overview of all LinB trajectories using the Chart Matrix view. They sliced the dataset using the *Minimum Active Site Distance* (rows) and *Area Under Curve (AUC)* (cols) properties, and observed the distribution of trajectories (see Figure 6). Then they brushed the data using the rectangular plot selection tool, in order to keep only those trajectories that reached the distance of 5 or less Å from the active site and were not outliers in terms of the total energy (*AUC*). This was performed in a few subsequent steps, since the binning became finer after each brushing step. Next, the biochemists switched the *AUC* property to the *Elevation* and removed outliers also according to this property. Finally, they created a selection from the remaining trajectories which represented a significantly large cluster of similar trajectories w.r.t. the binding energy.

After repeating the same process for the LinB32 dataset, they could explore two selections in the Selection Chart Panel. As they wanted to compare the trajectory selections in terms of overall energy, they represented them using functional boxplots, superimposed over each other. In the resulting plot, they could observe the energy peaks at the range of 8-14 Å from the active site in the LinB32 trajectories (see Figure 5). These peaks were noticeable on the median trajectory of the selection, as well as on the upper boundary of a region which represents the upper 50% of the selected trajectories when ordered according their closeness to the median.

From this observation, they concluded that even though the ligand was transported to the active site in some of the trajectories in LinB32, it was energetically less efficient. This case shows that our tool can easily compare trajectories from multiple datasets (**T6**) and provide the biochemists with the required information to draw important conclusions.

5.2 Conformation Changes of LinB

In the second study, the domain experts investigated the differences between the trajectories from the LinB dataset that transported the ligand to the active site. They started with the Chart Matrix slicing the data into 10 by 10 bins using the *Min Active Site Distance* (rows) and *Area Under Curve* (columns) properties (see Figure 6). As they were interested only in the trajectories that transported the ligand to the active site, they selected the data to drill down to the two bottom rows of the matrix. Then, they replaced the *Min Active Site Distance* by the *Energy Elevation* property for binning of the rows of the matrix. They further selected rectangular region of the most populated bins, making sure that the selection contained all bins with at least 100 trajectories. After that, they decreased the number of visualized bins to 6 by 6 (see Figure 2 in the Supplementary Material). In the two bottom rows of the matrix, it was possible to observe relatively flat trajectories that contained energy peaks only in the vicinity of the active site. The presence of these flat trajectories fades towards the top part of the view, where trajectories with the increasing value of *Energy Elevation* property are occurring. This is also confirmed by the vertical Marginal Charts on the left side of the matrix.

From the information derived so far, the domain experts suspected that the input MD simulation contained a conformation change of the protein which could cause the presence of these trajectories. To confirm or reject this hypothesis, they modified the binning, exchanging

the *Energy Elevation* with the *Structure RMSD* property. Note that in this case the RMSD could be computed for each trajectory as a scalar value because each trajectory corresponds to one and only one snapshot in the MD, due to the way the trajectories were simulated.

After this change, it was possible to observe the presence of similar bins containing the suspected trajectories – they emerged in the top part of the matrix which shows the bins with RMSD of 1.6-2.1 (see Figure 3 in the Supplementary Material). These RMSD values already proved the conformation changes w.r.t. the mean conformation of the protein. Nevertheless, to further investigate this, the biochemists visualized the trajectories and the protein in 3D. They used aggregated visualizations for both trajectories and protein (see Figure 3) which further strengthened the findings, regarding the conformation changes of LinB.

This demonstrates that our tool is easy to use for the exploration of large sets of functions describing physico-chemical properties along the trajectories (**T2**). When the biochemists tried to confirm these findings using their standard approach, they needed to measure the distance between two amino acids from the rigid and the moving parts of the protein. It was very time-consuming as they first needed to identify these amino acids. Moreover, they also claimed that the advantage of our tool is that it allows them to easily relate these conformation changes back to the ligand energy profiles (**T5**).

5.3 Feedback

After the biochemists conducted the case studies, we again arranged an informal interview with them in order to evaluate the design of our visualization system in detail. Overall, they all agreed that our solution helped them to perform the analysis of large trajectory datasets in an interactive manner and much faster than ever before. The biochemists found the Chart Matrix view to be the most informative and profitable representation applied their data. They confirmed that it enabled them to quickly examine their data and presented the trajectories in a clear and easily comprehensible manner. Both the density and boxplot aggregations of the energy data helped them to asses the main energy trends of highly similar trajectories already from the initial overview. Additionally, the selection-based drill-down functionality of the Chart Matrix helped them to reveal and analyze sparsely represented features of their data. The biochemists also highly appreciated the selection storing feature. They stated that it helped them to compare the similarities and relations of several subparts of one dataset as well as multiple datasets. They also often employed selections when they were examining fine details of their data. In regards to this task, they noted that it would be beneficial to be able to quickly revert the brushing they have just made, or, ideally, to be able to browse through the history of all selections and visualization changes. Furthermore, the biochemists often employed the interactive link between the Chart Matrix and the 3D View. It enabled them to investigate the spatial aspects of a subset of the data in question. Lastly, the biochemists showed interest in exploring the changes among the amino acids surrounding the ligand. This revealed one limitation of our current solution.

Our system already enables the biochemists to employ the minimal distance to the amino acids in the active site for binning purposes. This option can be easily extended for any selected amino acids. The system can also be extended to visualize the distance of ligand to the selected amino acid as a function along the trajectory instead of the binding energy. Finally, it is already possible to assess the surrounding amino acids from our 3D visualizations. However, the biochemists would rather explore the set of surrounding amino acids and the changes among them in a more straightforward way.

6 Conclusion and Future Work

In this paper, we proposed a novel visualization system which solves the problem of the analysis of ligand binding simulation data by conveying the representations of thousands of ligand trajectories. Our system enables the biochemists to analyze their data using a measure that is crucial for them – the *binding energy*, in an interactive and explorative manner. The system enables the users to explore their data while drawing conclusions from it in parallel. Thus none or only a minimal a priori knowledge is required with respect to the data analysis.

In order to verify the usefulness of our tool, the collaborating protein engineers conducted two case studies using two different datasets. The case studies confirmed that our system supports the tasks (**T1** - **T6**) that were elicited based on the needs of the domain experts and served as a main source in the design process of the tool. The goal of our tool was to support these tasks in order to make the system applicable to various problems related to analysis of ligand binding simulations, e.g., comparison of ligand binding to different proteins or binding of multiple ligands to the same protein.

In the future, we plan to extend the Chart Matrix visualization in order to allow the exploration of trends of trajectory properties with multidimensional ranges. Effectively, this would enable the biochemists to analyze even more complex aspects of their trajectory data (e.g., surrounding amino acids). Additionally, we plan to evaluate the physico-chemical properties of the surrounding amino acids as a part of our future trajectory visualization enhancements. Another possible future improvement would enhance the exploration of trajectories in the 3D view, which is still very cluttered. Therefore, introducing clipping planes or focus lens techniques would facilitate this task.

Finally, we believe that our system can be applied to other domains as well (e.g., traffic data or flow simulations) where the analysis of large trajectory data plays a significant role. Our proposed tool can be applied to any trajectory dataset for which the following two conditions are met: first, the domain-specific properties along the trajectory can be represented as simple scalar functions, and second, there are some meaningful properties that describe each trajectory with a single scalar value to enable the binning.

Acknowledgments

The presented work has been supported by the Czech Science Foundation (GAČR) under research project No. 17-07690S and by the MŠMT project no. LM2015055. Access to computing and storage facilities owned by parties and projects contributing to the National Grid Infrastructure MetaCentrum was provided under the programmes LM2010005, LM2015042, and LM2015085. O.Vávra is recipient of Ph.D. Talent award provided by the Brno City Municipality. We also would like to thank our collaborators at Loschmidt Laboratories, namely Dr. Piia Kokkonen for providing us with the datasets and Dr. Sérgio Marques for his invaluable feedback.

References

[1] N. Alharbi, R. S. Laramee, and M. Chavent. MolPathFinder: interactive multi-dimensional path filtering of molecular dynamics simulation data. In *The Computer Graphics and Visual Computing (CGVC) Conference 2016*, 2016.

[2] N. V. Andrienko and G. L. Andrienko. Spatial generalization and aggregation of massive movement data. *IEEE Transactions on Visualization and Computer Graphics*, 17(2):205–219, 2011.

[3] K. Bidmon, S. Grottel, F. Bös, J. Pleiss, and T. Ertl. Visual abstractions of solvent pathlines near protein cavities. In *Computer Graphics Forum*, vol. 27, pp. 935–942. Wiley Online Library, 2008.

[4] J. Brezovský, P. Babková, O. Degtjarik, A. Fořtová, A. Góra, I. Iermak, P. Řezáčová, P. Dvořák, I. Kutá-Smatanová, Z. Prokop, et al. Engineering a de novo transport tunnel. *ACS Catalysis*, 6(11):7597–7610, 2016.

[5] M. Chavent, T. Reddy, J. Goose, A. C. E. Dahl, J. E. Stone, B. Jobard, and M. S. Sansom. Methodologies for the analysis of instantaneous lipid diffusion in md simulations of large membrane systems. *Faraday discussions*, 169:455–475, 2014.

[6] J. Cortés, T. Siméon, V. Ruiz de Angulo, D. Guieysse, M. Remaud-Siméon, and V. Tran. A path planning approach for computing large-amplitude motions of flexible molecules. *Bioinformatics*, 21(suppl_1):i116–i125, 2005.

[7] U. Demšar and K. Virrantaus. Space–time density of trajectories: exploring spatio-temporal patterns in movement data. *International Journal of Geographical Information Science*, 24(10):1527–1542, 2010.

[8] D. Duran, P. Hermosilla, T. Ropinski, B. Kozlíková, À. Vinacua, and P.-P. Vázquez. Visualization of large molecular trajectories. *IEEE Transactions on Visualization and Computer Graphics*, 2018.

[9] F. Ferstl, K. Bürger, and R. Westermann. Streamline variability plots for characterizing the uncertainty in vector field ensembles. *IEEE Transactions on Visualization and Computer Graphics*, 22(1):767–776, 2016.

[10] J. Filipovič, O. Vávra, J. Plhák, D. Bednář, S. M. Marques, J. Brezovský, L. Matyska, and J. Damborský. CaverDock: A novel method for the fast analysis of ligand transport. *arXiv:1809.03453 [physics.bio-ph]*, 2018.

[11] K. Furmanová, M. Jarešová, J. Byška, A. Jurčík, J. Parulek, H. Hauser, and B. Kozlíková. Interactive exploration of ligand transportation through protein tunnels. *BMC Bioinformatics*, 18(2):22, 2017.

[12] M. J. Harvey, G. Giupponi, and G. D. Fabritiis. ACEMD: Accelerating biomolecular dynamics in the microsecond time scale. *Journal of Chemical Theory and Computation*, 5(6):1632–1639, 2009.

[13] P. Hermosilla, J. Estrada, V. Guallar, T. Ropinski, À. Vinacua, and P.-P. Vázquez. Physics-based visual characterization of molecular interaction forces. *IEEE Transactions on Visualization and Computer Graphics*, 23(1):731–740, 2017.

[14] R. J. Hyndman and H. L. Shang. Rainbow plots, bagplots, and boxplots for functional data. *Journal of Computational and Graphical Statistics*, 19(1):29–45, 2010.

[15] J. Jacques and C. Preda. Functional data clustering: a survey. *Advances in Data Analysis and Classification*, 8(3):231–255, 2014.

[16] P. Kokkonen, D. Bednář, V. Dočkalová, Z. Prokop, and J. Damborský. Conformational changes allow processing of bulky substrates by a haloalkane dehalogenase with a small and buried active site. *Journal of Biological Chemistry*, 293(29):11505–11512, 2018.

[17] Z. Konyha, K. Matković, D. Gračanin, M. Jelović, and H. Hauser. Interactive visual analysis of families of function graphs. *IEEE Transactions on Visualization and Computer Graphics*, 12(6):1373–1385, 2006.

[18] T. Koudeláková, R. Chaloupková, J. Brezovský, Z. Prokop, E. Šebestová, M. Hesseler, M. Khabiri, M. Plevaka, D. Kulik, I. Kutá-Smatanová, P. Řezáčová, R. Ettrich, U. T. Bornscheuer, and J. Damborský. Engineering enzyme stability and resistance to an organic cosolvent by modification of residues in the access tunnel. *Angewandte Chemie International Edition*, 52(7):1959–1963, 2013.

[19] B. Kozlíková, M. Krone, M. Falk, N. Lindow, M. Baaden, D. Baum, I. Viola, J. Parulek, and H.-C. Hege. Visualization of biomolecular structures: State of the art revisited. In *Computer Graphics Forum*, vol. 36, pp. 178–204. Wiley Online Library, 2017.

[20] M. Krone, B. Kozlíková, N. Lindow, M. Baaden, D. Baum, J. Parulek, H.-C. Hege, and I. Viola. Visual analysis of biomolecular cavities: State of the art. *Computer Graphics Forum*, 35(3):527–551.

[21] M. Krone, B. Kozlíková, N. Lindow, M. Baaden, D. Baum, J. Parulek, H.-C. Hege, and I. Viola. Visual analysis of biomolecular cavities: state of the art. In *Computer Graphics Forum*, vol. 35, pp. 527–551. Wiley Online Library, 2016.

[22] R. Laskowski and M. Swindells. Ligplot+: multiple ligand-protein interaction diagrams for drug discovery. *Journal of Chemical Information and Modeling*, 51(10):2778–2786, 2011.

[23] T. Magdziarz, K. Mitusińska, S. Gołdowska, A. Płuciennik, M. Stolarczyk, M. Ługowska, and A. Góra. Aqua-duct: a ligands tracking tool. *Bioinformatics*, 33(13):2045–2046, 2017.

[24] M. Mirzargar, R. T. Whitaker, and R. M. Kirby. Curve boxplot: Generalization of boxplot for ensembles of curves. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):2654–2663, 2014.

[25] S. Pall, M. J. Abraham, C. Kutzner, B. Hess, and E. Lindahl. Tackling exascale software challenges in molecular dynamics simulations with gromacs. In *International Conference on Exascale Applications and Software*, pp. 3–27. Springer, 2014.

[26] M. Pavlová, M. Klvaňa, Z. Prokop, R. Chaloupková, P. Banáš, M. Otyepka, R. C. Wade, M. Tsuda, Y. Nagata, and J. Damborský. Redesigning dehalogenase access tunnels as a strategy for degrading an anthropogenic substrate. *Nature Chemical Biology*, 5(10):727–733, 2009.

[27] H. Piringer, R. Kosara, and H. Hauser. Interactive focus+context visualization with linked 2D/3D scatterplots. In *Coordinated and Multiple Views in Exploratory Visualization, 2004. Proceedings. Second International Conference on*, pp. 49–60. IEEE, 2004.

[28] A. Pobitzer, A. Lež, K. Matković, and H. Hauser. A statistics-based dimension reduction of the space of path line attributes for interactive visual flow analysis. In *Proceedings of IEEE Pacific Visualization Symposium*. IEEE, 2012.

[29] Z. Prokop, A. Góra, J. Brezovský, R. Chaloupková, V. Štěpánková, and J. Damborský. Engineering of protein tunnels: Keyhole-lock-key model for catalysis by the enzymes with buried active sites. *Protein Engineering Handbook*, 3:421–464, 2012.

[30] J. S. Richardson. The anatomy and taxonomy of protein structure. In *Advances in Protein Chemistry*, vol. 34, pp. 167–339. Elsevier, 1981.

[31] R. Skånberg, M. Linares, C. König, P. Norman, D. Jönsson, I. Hotz, and A. Ynnerman. VIA-MD: Visual Interactive Analysis of Molecular Dynamics. In J. Byska, M. Krone, and B. Sommer, eds., *Workshop on Molecular Graphics and Visual Analysis of Molecular Data*, pp. 19–27. The Eurographics Association, 2018. doi: 10.2312/molva.20181102

[32] Y. Sun and M. G. Genton. Functional boxplots. *Journal of Computational and Graphical Statistics*, 20(2):316–334, 2011.

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

[34] E. Tufte and P. Graves-Morris. *The visual display of quantitative information.* CT: Graphics Press, 1983.

[35] V. Vad, J. Byška, A. Jurčík, I. Viola, E. M. Gröller, H. Hauser, S. M. Marques, J. Damborský, and B. Kozlíková. Watergate: Visual exploration of water trajectories in protein dynamics. In A. H. S. Bruckner and B. Kainz, eds., *Eurographics Workshop on Visual Computing for Biology and Medicine*, pp. 33–42. Eurographics Workshop on Visual Computing for Biology and Medicine, Bremen, Germany, 2017.

[36] V. Vad, D. Cedrim, W. Busch, P. Filzmoser, and I. Viola. Generalized box-plot for root growth ensembles. *BMC Bioinformatics*, 18(2):65, 2017.

[37] S. van den Elzen and J. J. van Wijk. Small multiples, large singles: A new approach for visual data exploration. In *Computer Graphics Forum*, vol. 32, pp. 191–200. Wiley Online Library, 2013.

[38] S. Vassiliev, P. Comte, A. Mahboob, and D. Bruce. Tracking the flow of water through photosystem II using molecular dynamics and streamline tracing. *Biochemistry*, 49(9):1873–1881, 2010.

[39] P. Vázquez, P. Hermosilla, V. Guallar, J. Estrada, and À. Vinacua. Visual analysis of protein-ligand interactions. *Computer Graphics Forum*, 37(3):391–402, 2018.