Eurographics Workshop on Visual Computing for Biology and Medicine (2017)
S. Bruckner, A. Hennemuth, and B. Kainz (Editors)

Watergate: Visual Exploration of Water Trajectories in Protein Dynamics

V. Vad[1†], J. Byška[2†], A. Jurčík[3], I. Viola[1], E. M. Gröller[1], H. Hauser[2], S. M. Marques[3,4], J. Damborský[3,4], and B. Kozlíková[3]

[1]TU Wien, Austria
[2]University of Bergen, Norway
[3]Masaryk University, Czech Republic
[4]International Clinical Research Center, St. Anne's University Hospital, Czech Republic

Abstract

The function of proteins is tightly related to their interactions with other molecules. The study of such interactions often requires to track the molecules that enter or exit specific regions of the proteins. This is investigated with molecular dynamics simulations, producing the trajectories of thousands of water molecules during hundreds of thousands of time steps. To ease the exploration of such rich spatio-temporal data, we propose a novel workflow for the analysis and visualization of large sets of water-molecule trajectories. Our solution consists of a set of visualization techniques, which help biochemists to classify, cluster, and filter the trajectories and to explore the properties and behavior of selected subsets in detail. Initially, we use an interactive histogram and a time-line visualization to give an overview of all water trajectories and select the interesting ones for further investigation. Further, we depict clusters of trajectories in a novel 2D representation illustrating the flows of water molecules. These views are interactively linked with a 3D representation where we show individual paths, including their simplification, as well as extracted statistical information displayed by isosurfaces. The proposed solution has been designed in tight collaboration with experts to support specific tasks in their scientific workflows. They also conducted several case studies to evaluate the usability and effectiveness of our new solution with respect to their research scenarios. These confirmed that our proposed solution helps in analyzing water trajectories and in extracting the essential information out of the large amount of input data.

1. Introduction

Protein structures and their function have been intensively studied for decades, due to their importance in biochemical research where the interaction of proteins with other molecules plays a crucial role. Among these molecules we find substrates, products, cofactors, drugs, and water molecules. If the reaction site, i.e., the active site, is buried deeply in the protein, the presence of entrance paths, called tunnels, leading from the outer surface to this site is crucial. The characteristics of these paths, such as their width, length, curvature, or physico-chemical properties of the surrounding amino acids and their changes over time, can directly influence the protein properties and function [LvB*17]. Similar properties have to be studied when examining the use of different paths by several molecules. In the case of a ligand passage, the biochemists typically deal with the trajectory of one or few ligands and try to understand their behavior over time. When taking into account the flow of water molecules through the tunnels, however, the task extends to analyzing hundreds to thousands of trajectories. These trajectories are strongly scattered because of the Brownian motion of water molecules (see Figure 1). It is very difficult to understand the behavior of so many water molecules without advanced visualization.

The importance to study the flow of water molecules through proteins has been demonstrated, for example, by the case study of Pavlova et al. [PKC*09]. The authors aimed to redesign the wild type of a haloalkane dehalogenase molecule in order to increase its reactivity with the toxic pollutant 1,2,3-trichloropropane (TCP). They proposed that the low reactivity of the wild type might be caused by the presence of many water molecules flowing to the active site through the main entrance tunnel. This required analyzing long simulations of molecular dynamics and observing the behavior of hundreds of water molecules. Without a proper visual support, this task is very cumbersome and complex to perform.

Another example to highlight the importance of studying water flow through protein cores is the assumption that the flow may induce an opening of a molecular path. This presumes that if a tunnel is used by water molecules flowing towards the active site, this tunnel might be the relevant path for ligands and other small molecules as well, and hence play a biological role. However, this hypothesis was not confirmed yet, so with our new visualization we are aiming to help the biochemists to understand this problem better.

[†] These authors contributed equally to this work.

DOI: 10.2312/vcbm.20171235

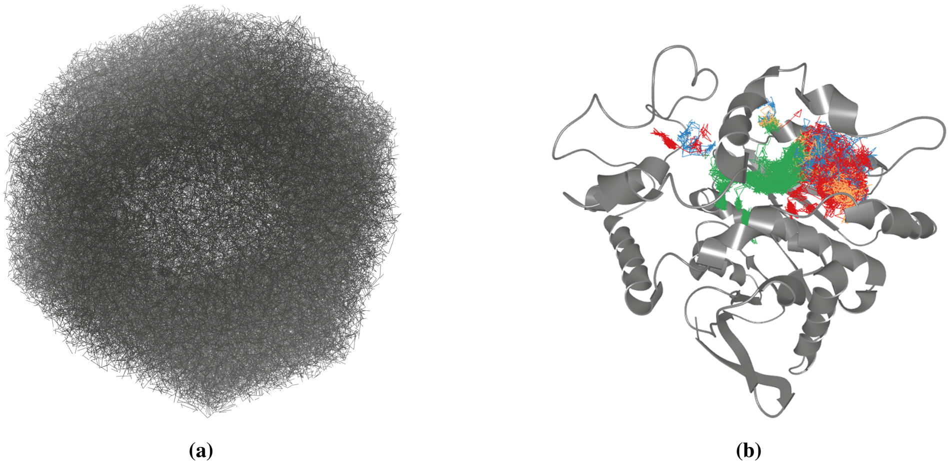


Figure 1: *(a) More than 11.000 water trajectories over only 25 time steps (from 100.000 time steps of the whole simulation). (b) Colored trajectories of 1.077 water molecules that actually reached the active site during the simulation.*

The above assumptions and requirements led us to the design and development of a set of visualization techniques for the interactive exploration and analysis of thousands of water trajectories within long simulation sequences. Our interactive, linked visual representations aim at helping the biochemists to reveal the most relevant water flows and to analyze their overall characteristics as well as details of individual behavior. Our solution starts with a basic analysis of all trajectories of water molecules with respect to their penetration through the protein surface and entering the active site region. This divides the initial set of trajectories into several categories. The user can then visually compare these categories and select a subset for further exploration, which is performed using linked 2D and 3D views. The contributions of this work are:

- A novel visualization of clusters of water molecules, which helps to understand their numbers and flow directions.
- A visual analytics framework for the interactive visual analysis of water molecules. It applies a set of traditional as well as novel visual representations supporting the workflow of protein engineers.

2. Related Work

The visualization and visual analysis of biochemical data is a vast area, which has been in the focus of researchers already for decades. Recent surveys on the visualization of biomolecules as well as an analysis and visualization of their cavities by Kozlíková et al. [KKF*16] and Krone et al. [KKL*16] confirm the importance of visualization in biochemical and biological research.

2.1. Simplification of Trajectories and their Visualization

As already stated, water trajectories inside proteins are strongly scattered because of Brownian motion. Similar trajectories, i.e., scattered and time-dependent, can be found in many other fields as well. Therefore, there already exist many solutions for trajectory simplification.To find the most suitable simplification method, Dodge et al. [DWL08] introduce a taxonomy of different movement patterns. They categorize types of motion and propose an appropriate solution for their analysis and visualization. Another valuable source of existing tools and techniques comes from the book by Andrienko et al. [AAB*13]. Also, recent work of Vrotsou et al. [VJN*15] presents a systematic stepwise methodology for trajectory simplification with an emphasis on visual exploration and analysis. Trajectory simplification has been also used in the work of Furmanová et al. [FJB*17]. The authors are focusing on the exploration and visual analysis of ligand trajectories using various linked views. Here the proposed methods are tailored to explore individual ligand trajectories and therefore cannot be used for studying trends of large flows of water molecules and their filtering. Luboschik et al. [LRB*15] investigate the dependency of synthesized movements on parameters that control the simulation generating these movements. The authors use visual analytics approach to extract features that describe the distinct movements in order to simplify the complexity of the problem.

In our case the biochemists are not interested in individual trajectories but rather in a statistical overview of the whole flow of water molecules. Holten [Hol06] introduce edge bundling, a visualization technique for dense graphs. The author groups adjacent edges in order to reduce cluttering in the visualization. Phan et al. [PXY*05] use hierarchical clustering to group nodes in a way that edge crossings were minimized, while the relative positions of nodes were preserved. Andrienko et al. [AA11] extract significant points from trajectory data. Then, they use arrows, which are pointing from/to certain areas represented as cells of Voronoi diagrams. These areas are used for the representation of movement trends. Cornel et al. [CKS*16] introduce composite flow maps, which enable to visualize multiple flows at the same time. Space-time movements can also be modeled as probability densities. Demsar et al. [DV10] use three dimensional kernels (2D for space, 1D for time) to model space-time movement data. Stegmaier et al. [SRE05] focus on understanding of turbulent flows where they use vortices to visualize the dynamics. As turbulence is not an important aspect of water molecules, this technique cannot be adapted to our case.

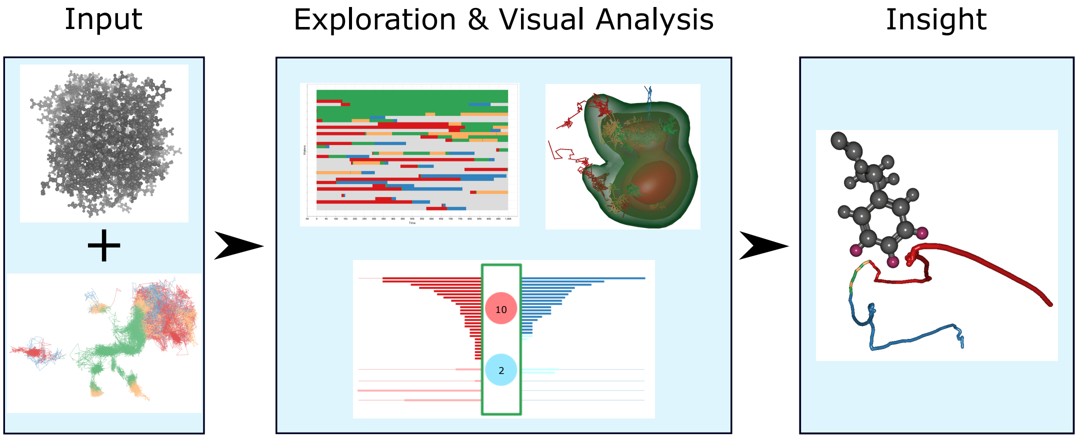


Figure 2: *System overview. Input data (simulation of molecular dynamics of a protein and trajectories of water molecules flowing through this protein) are processed by our exploration and visual analysis tool. The results can be examined in the 3D view.*

2.2. Analysis and Visualization of Water Trajectories

Water trajectories and the importance of their analysis has been already addressed by several research groups. Benson and Pleiss [BP14] describe a method for routing the water flow in a given direction. Very recently, the AQUA-DUCT [MMG*16, MMG*17] tool was released. It allows the user to extract, analyze, and visualize the behavior of solvent molecules during a molecular dynamics simulation. The visualization is currently supported by the PyMol visualization package [Sch15].

In addition to basically visualizing the shape of individual trajectories, there are also techniques addressing the understanding of the underlying flow. Among these techniques is the work of Vassiliev et al. [VCMB10], where the authors visualize the water trajectories using streamlines.

The work of Bidmon et al. [BGB*08] is closely related to our approach. It focuses on the identification and visualization of water trajectories within molecular dynamics. The authors propose to cluster trajectories, followed by a 3D visual abstraction, which shows only the extracted principal paths. Their approach preserves valuable information on the directions and velocities of water molecules, which are moving along these paths. However, their tool does not provide any sophisticated interaction with the visualized trajectories. For instance, the informed selection and exploration of an interesting subset of trajectories (e.g., trajectories entering through a single molecular tunnel) is not possible. Luboschik et al. [LMS*12] present another approach that describes a guidance through complex simulation trajectories in systems biology. Ertl et al. [EKK*14] propose a visual analysis tool for space-time aggregation of DNA simulations, which uses vector fields. Similarly, Chavent et al. [CRG*14] use path lines, vector fields, and streamline visualization for the visual analysis of molecular positions in dynamics simulations and their velocity fields. Recently, Alharbi et al. [ALC16] presented a tool for multi-dimensional path filtering of molecular-dynamics simulation-data. It consists of a set of functions to interactively filter and highlight dynamic and complex paths from the motion of molecules, which are visualized as curves.

3. Watergate Overview

The input data consists of molecular dynamics simulation of a protein, containing usually thousands of complex trajectories of water molecules. Additionally, the spatial position of the protein active site is specified by the domain experts, derived from a set of amino acids. The goal of the biochemists is to understand the main trends in the trajectory behavior. This can be expressed by the following requirements and questions that were posed directly by the domain experts.

- The biochemists want to observe only those water molecules that are entering the protein.
- For these trajectories of water molecules, the biochemists want to get information about the behavior – how and when did the water molecules get to the active site, how long did they stay in the active site, did they leave the protein using the same tunnel or did they use another one, etc.
- Which tunnels were used for entering and leaving the protein and to which extent? How much the known tunnels overlap with the space occupied by water molecules during the molecular dynamics simulation?

To address these issues and questions, we propose a system consisting of several analysis and visualization steps, which guide the biochemist through the process (see Figure 2). All proposed visualizations are interactively linked. If the user selects a set of trajectories using a given visualization, the corresponding trajectories are highlighted in the remaining views as well. There are also several specific interaction techniques, which are further described along with the associated visualizations.

The process starts with the initial set of all trajectories, containing the information about the positions of thousands of water molecules over thousands of time steps of the molecular dynamics simulation. First, we analyze all trajectories with respect to their entrance to the protein. If a water molecule never enters the protein, it is automatically discarded in this step. This immediately fulfills the first requirement stated by the domain experts. The remaining trajectories are divided into segments according to a classification used by the AQUA-DUCT [MMG*16] tool.

3.1. Classification of Trajectories

Figure 3 illustrates the classification of trajectories according to the spatial position of corresponding water molecules over time. First, we divide the space of the protein into three main regions according to importance. One of the most interesting parts of the protein is the active site region (marked by the green curve in Figure 3). We call it the **focus region**. The rest of the inner space of the protein defines the second region, called the **inner region**. The third region, the **outer region**, is formed by the outer environment (i.e., the surroundings of the protein). The boundary between the inner and outer region is determined by the protein surface. In our application we represent the protein surface through an isosurface of a Gaussian density field [KSES12] with the isovalue set to 1.5, which nicely corresponds to the protein surface shape.

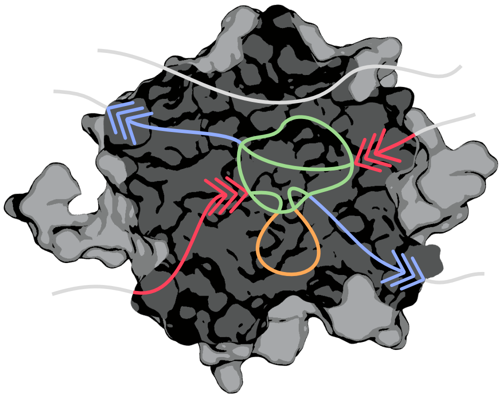


Figure 3: *Illustration of the trajectory classification. The protein is represented by a surface and the curves illustrate individual segments of trajectories.*

For each trajectory, we analyze the position of the water molecule in each time step with respect to these predefined regions. This divides the trajectory into several segments. Additionally, in the inner region, we also calculate the direction vector of the water molecule. This determines if the water molecule is heading towards the focus region or away from it. The molecule can also travel inside the inner region without principal direction and can reenter the focus region several times. According to these categories, the segments of trajectories can be marked as:

- **incoming** – enters the inner region at the protein surface, is heading towards the focus region, and at some point enters the focus region (red curves in Figure 3),
- **outgoing** – leaves the focus region, is heading towards the protein surface, and finally leaves the inner region (blue curves in Figure 3),
- **inside** – is in the focus region or in its user-defined vicinity (green curves in Figure 3),
- **around** – is in the inner region, does not follow any of the two main directions, and reenters the focus region during the simulation (orange curve in Figure 3),
- **outside** – is in the outer or in the inner region but is never getting to the focus region (grey curves in Figure 3).

This classification is used by the proposed visualization techniques to color and filter the water trajectories. To be able to detect the main flow directions, i.e., to detect the tunnels used by the water molecules, we further process the trajectories using clustering.

We cluster points in 3D space, which are intersections of individual trajectories with the protein surface. These points represent entrance/exit sites of water molecules on their way to/from the protein. By clustering them, we get regions on the protein surface, which should in fact correspond to tunnel entrances.

During the development we found out that this clustering could be achieved in a non-linear 2-fold space using a geodesic distance on the protein surface. Since we are dealing with molecular dynamics simulations where the molecular structure is changing over time, we have to compute the corresponding Gaussian surface in each time step to check if a trajectory interacts with the protein.

In our application we use the mean shift clustering algorithm [FH75] for 3D point clouds. Mean shift clustering is an iterative method to identify local maxima (modes) in an estimated density function. One of the method's advantages is that it does not need any a priori information about the number of clusters, unlike other popular clustering methods (e.g., K-means). This is an important issue in our application, since in this phase it is not known how many entrances (clusters) are present in the dataset. The resulting clusters serve as an input for the subsequent visualizations, which give an overview of the trajectories.

3.2. TimeLine View

The first proposed visualization conveys the classification of individual trajectories. It is based on time-line visualizations, mainly the color line view [MGKH07]. Each row depicts one water trajectory and the abscissa corresponds to time (see Figure 4(a)). The coloring of segments of each row corresponds to the classification.

Figure 4(b) shows that the colors in some rows change dramatically. This signifies that the water molecules are often changing their position according to the classification. If the biochemists are interested only in the most significant changes along the trajectory,

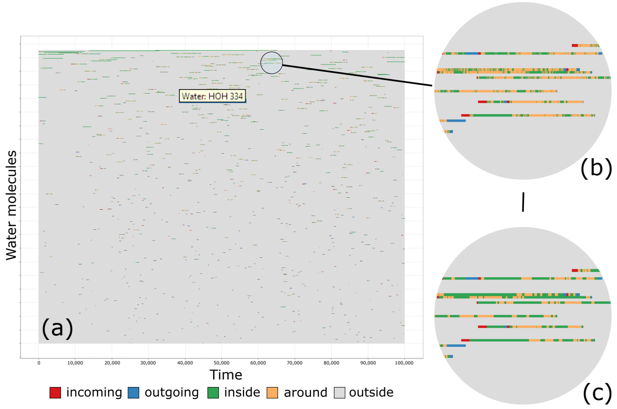
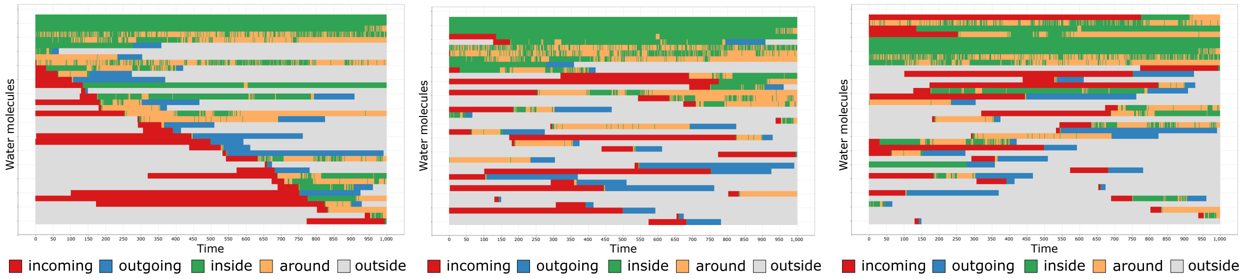


Figure 4: *TimeLine View of 1.077 input trajectories over 100.000 time steps (a), with close-up views (b), (c).*



(a) **(b)** **(c)**

Figure 5: *TimeLine View of a small dataset consisting of 37 water molecules, vertically sorted according to different criteria: (a) Time when the water molecule enters the focus region for the first time. (b) Overall time the water molecule spent in the focus region. (c) Overall time the water molecule spent inside the protein.*

this information is too detailed. Therefore, we simplify fast changes occurring close to the boundary of the focus region. These changes are visualized as a set of thin green segments alternating with thin orange ones. They are now classified as being inside the focus region. We apply the following set of simple rules. The segment between the point, where the water enters the protein and the point, where it enters the focus region for the first time is marked as **incoming**. The segment between the point, where the water finally leaves the focus region and the point, where it leaves the protein is marked as **outgoing**. We mark the part of the trajectory as **around** if and only if the water molecule leaves the focus region for a substantial portion of time (according to a user-defined parameter), otherwise it is marked as **inside**. This results in a more compact representation as can be seen in Figure 4(c). This also supports the scalability of the proposed view because in case of very long molecular dynamics simulations the high frequency details are not visible anymore. Nevertheless, the user can easily switch between both representations when needed.

In both representations the water molecules can be sorted and filtered along the vertical axis according to the following criteria (see Figure 5):

- time when the water enters the focus region for the first time,
- overall time spent in the focus region,
- overall time spent inside the protein.

The TimeLine View cannot natively handle very large datasets. In case of long molecular dynamics simulations, the above mentioned sorting and filtering helps to narrow down the dataset to only its most important parts. The user can also interact with the time-line visualization in other ways. The user can zoom-in to explore a specific region of the TimeLine View in more detail. While zoomed-in, the user can pan in horizontal and vertical direction to adjust the view as needed. For better user comfort, we have implemented an auto-range function that computes a bounding box enclosing all points in the chart. It sets zoom and panning in such a way that all the data are clearly visible. The TimeLine View also allows the user to select a specific area of interest along both the abscissa and the ordinate to mark a specific combination of time steps and water molecules. In this way the user can filter also the selected time range. These selections are automatically visualized in the subsequent proposed views.

The individual labels on the ordinate of the TimeLine View could easily become very small or start to overlap, so we decided to leave them away completely. Instead, we have implemented a tooltip window (see Figure 4(a)). It appears when the user hovers over a data point in the TimeLine View, providing the information about the exact corresponding water molecule ID. The TimeLine View provides a general information about the behavior of water molecules. It also explicitly shows the time span for which a water molecule belonged to a given category and hence answers the second research question posed by the domain experts.

3.3. WaterFlow Map

The main purpose of the WaterFlow Map is to show the behavior of water molecules with respect to the path they take through the protein. This information helps the biochemists to understand the importance of individual tunnels and their role in the protein function. The clusters described in Section 3.1 explicitly divide the trajectories into groups according to the parts on the surface they took to enter/exit the protein. Hence, the flow map simply visualizes the overview of all detected clusters, along with their sizes, time evolution, and classification.

Figure 6 depicts a WaterFlow Map. It consists of the central (green) rectangle representing the focus region. On the left and right side of this rectangle, there are lines corresponding to individual incoming (left) and outgoing (right) water molecules with

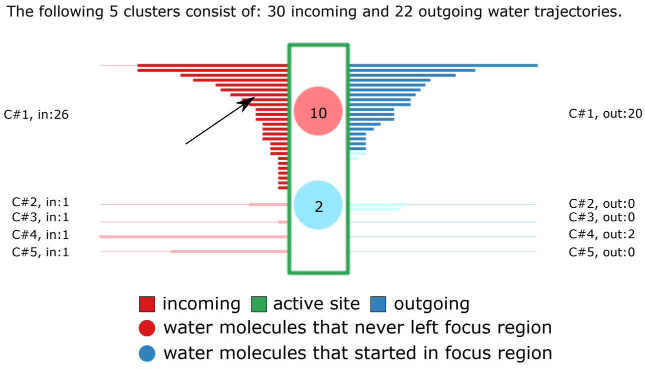


Figure 6: *WaterFlow Map showing the incoming and outgoing clusters of water trajectories.*

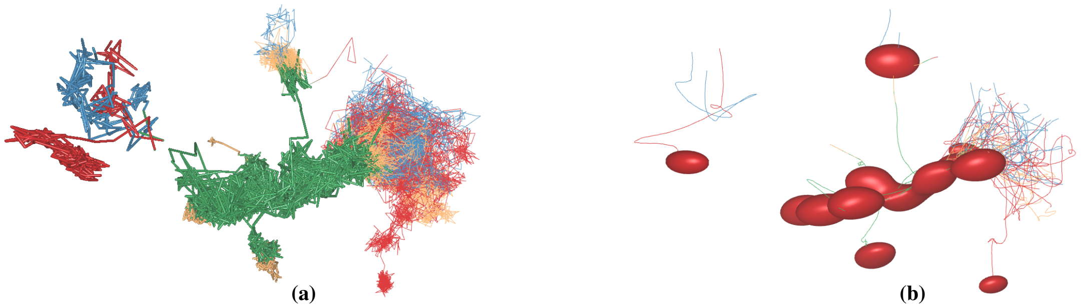


Figure 7: *(a) Combination of line and tube representations of trajectories. (b) Simplified version of these trajectories using the Savitzky-Golay algorithm. Red surfaces define areas where the water molecules stayed for a significant portion of time.*

respect to the focus region. Line length corresponds to the time span it takes the corresponding water molecule to get to/from the focus region.

Contrary to the TimeLine View, where each line corresponds to a single water molecule, in the WaterFlow Map each line corresponds only to a segment of the trajectory. This means that some water molecules are depicted multiple times if they entered the protein multiple times. This design choice better supports the workflow of biochemists who are not interested in individual water molecules but rather in a statistical overview of the system. In such a case, a single water molecule entering the protein twice has the same impact as two water molecules, each entering the protein just once.

Next to each cluster we also show the exact number of member trajectories. This allows the user to see the overall number of water molecules entering or leaving through a particular cluster. The green rectangle, representing the focus, contains the information about the exact number of two types water molecules. Those, which entered the focus region and stayed there (red circle) and those, which were already in the focus region at the beginning of the simulation and left the protein inner region (blue circle).

The WaterFlow Map supports two levels of interaction. The user can hover over a single water trajectory or a whole cluster, e.g., the large incoming one in Figure 6 indicated by the black arrow. The visualization then automatically highlights the corresponding outgoing trajectory or multiple trajectories in case of cluster selection on the right side of this view (dark blue color). The user can immediately see how these trajectories behaved later on, i.e., how many of them stayed in the focus region, how many of them left, and which clusters did they use.

As a given cluster can be used both as incoming and outgoing route in the simulation, the corresponding incoming and outgoing lines of this cluster are always located at the same horizontal position (on the opposite side of the green rectangle of the focus region). If a given cluster is used only as incoming, the corresponding outgoing cluster is empty, which is depicted by a very thin line in the WaterFlow Map.

When hovering over the clusters, the corresponding trajectories are automatically highlighted in the other proposed views. The user can also manually select more than one cluster of interest and proceed with a further analysis. The WaterFlow Map hence directly fulfills the third and last requirement from the biochemists. Figure 9 demonstrates that our representation is applicable to large datasets as well. It shows the extracted clusters for a simulation containing more than 11.000 water molecules. More than 1.000 are depicted in the image, i.e., those that reached or left the focus region. For large data sets, the WaterFlow Map automatically aggregates trajectories from the same cluster with similar length to show the big picture of all trajectories and their clustering. In such a case, each line corresponds to several trajectories. The algorithm estimates the available space and sets the aggregation factor (i.e., the number of trajectories represented by a single line) in such a way that the resulting visualization fits to the screen.

3.4. 3D Views – Trajectories and DensityIsosurface

In the three-dimensional space, we provide the user with several representations of selected trajectories on different levels of visual abstraction. The user can visualize the original trajectories (Figure 7(a)) or their simplified versions (Figure 7(b)).

The original or simplified trajectories can be subsequently visualized as simple lines or as tubes (see Figure 7(a)). Tubes help to better understand the spatial arrangement of trajectories and are used for highlighting parts of trajectories selected via the previously described 2D views. To preserve the context information, the user can additionally visualize the remaining parts of the trajectories as lines. Both tubes and lines can be colored according to different criteria, such as time, detected clusters, or classification. The amount of displayed trajectories can be also reduced through filtering in some of the previously described 2D views.

For the trajectory simplification, we apply the Savitzky-Golay [SG64] filter. This method is widely used, for instance, in physics or chemistry to increase the signal-to-noise ratio without altering the original data significantly. It produces smooth curves (see Figure 7(b)) that help to communicate the main trends in the shape of water trajectories. The main disadvantage of this approach is that the simplification shifts the original positions of water molecules obtained from the molecular dynamics simulation. To address this issue, we experimented with several non-smoothing simplifications, for instance, the Douglas-Peucker [VW90] algorithm.

One of the problems of all tested simplification methods is that they remove parts of the trajectory where the water molecules spent a significant amount of time. From a biochemical point of view, this information should be preserved because it signifies possibly interesting sites. Therefore, in addition to the direct trajectory visualization, we propose an additional visualization (Figures 7(b) and 8). It aims to summarize the occupancy information of the protein inner region with selected water molecules over time. The goal is to convey the information where in space the molecules from trajectories of interest were most likely present in a specific time span. This view is very important because the biochemists can instantly get the summary information about the protein inner space occupied by water molecules.

The visualization is based on estimating the probability density function in three dimensional space. Then, isosurfaces representing different percentiles of the estimated density function are shown. In Figure 7(b) only the 25% percentile is depicted. Here we were inspired by Raunest and Kandt, who developed dxTuber [RK11]. It is a software tool for detection of cavities, tunnels, and clefts in protein molecules coming from dynamic simulations. Their tool uses a discretized volume representation, where each voxel contains a ligand and the protein mass density for a time period. However, their approach does not show the probability of ligand occupancy, which is what the biochemists require in our case.

We use kernel density estimation (KDE), which is a non-parametric density function estimation. It estimates a probability density function (PDF) based on discrete input samples. In 3D, the estimation is formulated as

$$f(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} K_{\mathbf{H}}(\mathbf{x}-\mathbf{x}_i), \tag{1}$$

where $\mathbf{x}_i, i = 1 \dots n$ are the input sample positions, K is a 3D kernel function, $\mathbf{H}$ is a symmetric positive definite bandwidth matrix that determines the size and shape of the kernels in the estimation, and the scaled kernel is defined as

$$K_{\mathbf{H}}(\mathbf{x}) = |\mathbf{H}|^{-1/2} K\left(\mathbf{H}^{-1/2}\mathbf{x}\right). \tag{2}$$

where $|\mathbf{H}|$ is a determinant. Many different kernels K are used in practice. However, most of the times, and in our application as well, a Gaussian kernel is chosen. In order to increase the efficiency of the density estimation, we estimate and use an unconstrained, symmetric, positive definite bandwidth matrix $\mathbf{H}$. Therefore, our scaled kernels are anisotropic and not axis-aligned. Whenever the selection of trajectories changes, the input sample positions can also change dramatically. Therefore, the underlying density changes as well. So for each trajectory and time span selection we estimate a different bandwidth matrix. The direct evaluation of f in Equation 1 can be computationally expensive, especially in such a scenario where n is large. In order to react to the selection of trajectories in an interactive manner, we use parallel evaluation of f and efficient state-of-the-art bandwidth matrix estimation. More details are provided in the Supplementary Material.

Equation 1 defines the density for the entire 3D space. Our intention is to visualize the estimated density in an expressive way. Hence, we depict percentile volumes, which are generalizations of the univariate case, as DensityIsosurface representation. A p^{th} percentile of the probability density field corresponds to an isosurface, which encapsulates those parts of the domain having the highest density values. Moreover, the integration of the encapsulated part of the probability density field results in p percent.

In order to render the isosurface of a given percentile, we have to find the isovalue that corresponds to this percentile. In the univariate case it is done by inverting the cumulative density function. However, in our multivariate case and for general PDFs computing the inverse function is not tractable, so we need an approximation. We define a regular volumetric grid, and at each voxel, we evaluate function $f(\mathbf{x})$ of Equation 1. The size of the grid corresponds to the bounding box of the molecule. The samples $\mathbf{x}_i$ from Equation 1 correspond to individual positions of selected water molecules over a selected time span.

After we evaluate f for all voxels, we sort the obtained discrete voxel values in a descending order, multiply them with the voxel volume, and sum up. If we would sum up all the values, we would end up at one, since f is a probability density function. We stop the summation as soon as we reach the desired percentage p. The last density value used in the summation is the isovalue of the specified percentile.

For the surface extraction, we use the Marching Cubes algorithm [LC87], and the surface is rendered as a triangle mesh. Examples of percentile isosurfaces can be seen in Figure 8. The 25% percentile isosurface (red) shows the part of the protein which is most densely occupied by water molecules. This region is located around the protein active site. Therefore, this representation could be potentially used also for estimating the position of active sites for yet unknown proteins.

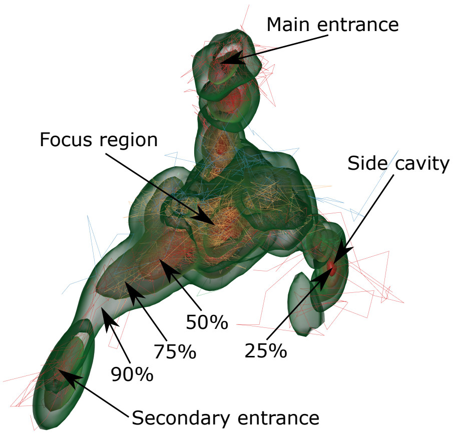


Figure 8: *DensityIsosurface layers representing the 25%, 50%, 75%, and 90% percentiles for a dataset containing two clusters. The isosurfaces are combined with line representations of the corresponding trajectories, colored according to their classification. Two entrance paths (corresponding to the two identified clusters) were discovered.*

When evaluating the performance of our solution, we measured the elapsed time for computing the binned grid, estimating the bandwidth matrix, estimating the density volume, estimating the percentile isovalues, and managing the GPU resources. In case of about 10.000 points, the computations needed around 800 ms in a grid of resolution 128^3. The Supplementary Material contains the table comparing the performance of our solution with the existing ones.

All proposed representations can also be combined with tunnels detected by the CAVER tool [CPB*12] (see Figure 5(b) in the Supplementary Material). In consequence, this can help to answer the hypothesis about the biochemical relevance of tunnels and flows of water molecules. This hypothesis is based on the correlation between the importance of a tunnel and its usage by water molecules.

4. Case Study & Discussion

To demonstrate the use of Watergate, we present an informal case study performed by protein engineering experts on data from their own research project [PKC*09]. However, we tested our solution on many other cases as well. Another example can be found in the Supplementary Material. The presented case study aims to show that all three requirements defined by domain experts and described at the beginning of Section 3 were fulfilled. We analyzed a molecular dynamics simulation of a variant of the haloalkane dehalogenase LinB in a box of an explicit solvent. This simulation contains the trajectories of the system during a time period of 200 ns and is composed by 100.000 time steps spaced by 2 ps, and contains more than 11.000 trajectories of water molecules.

First, all trajectories were analyzed with respect to their interaction with the protein. Those water molecules that never enter the protein were automatically discarded. The remaining 1.077 water molecules were classified according to the proposed categorization and the clusters of incoming and outgoing trajectories were detected. These steps were performed in a preprocessing stage and served as an input for the subsequent visualizations. This part of the analysis fulfills the first requirement to *"observe only the water molecules that are entering the protein"*. This was already possible to reach before, e.g., with the AQUA-DUCT [MMG*17] tool, which uses the convex hull for determining the outer region of the protein. But in our approach, we use the protein surface representation using isosurface which is more precise than the convex hull.

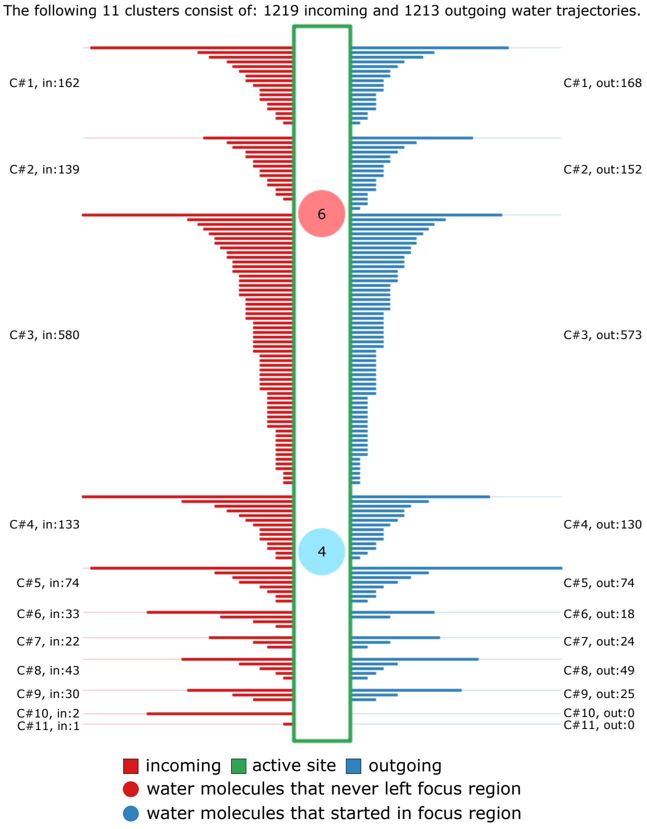


Figure 9: *WaterFlow Map used in the case study.*

In order to answer the question, how many water molecules reached the active site, the WaterFlow Map can be used (see Figure 9). In this particular case study WaterFlow Map revealed that the water molecules were divided into 11 clusters and one of the clusters was highly dominant, containing over 580 water trajectories. Furthermore, the distribution of trajectories between incoming and outgoing trajectories was rather uniform.

To analyze how much time each water molecule spent in the active site, the biochemists used the TimeLine View and sorted the trajectories according to retention time (see Figure 4). This view immediately communicated the most important water molecules for further evaluation. With the TimeLine View the biochemists concluded that only a small portion of water molecules was actually in the vicinity of the active site for a significant number of time steps during the simulation. These molecules are depicted as long green lines. This already provided relevant information on the high dynamics of the simulated system. In the next step all water molecules that were present in the active site for less than 500 time steps were filtered out, in order to preserve only the 224 most important water molecules. The threshold and hence the number of selected water molecules can be changed on the fly by the user. Figure 10 shows the selected 224 water molecules, sorted according to the time they reached the active site for the first time. As can

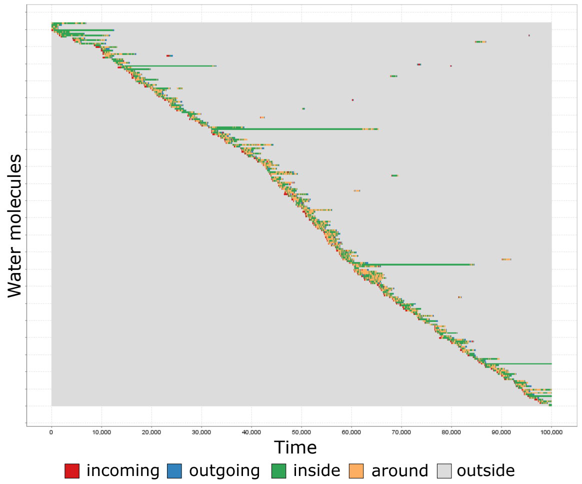


Figure 10: *TimeLine View showing the trajectories of 224 water molecules. Long green lines depict water molecules, which stayed in the active site for a long time.*

be seen, the water molecules are regularly distributed along the diagonal. This confirms that only a small number of water molecules is present in the active site at the same time. This information was crucial, since it demonstrated that the hydration of the active site had reached an equilibrium. It also indicated that the binding of a ligand to the active site would not be obstructed by many water molecules being there. This revealed to be a problem in some cases described in [PKC*09].

In the next step, the main trajectories used by the water molecules to reach the active site were analyzed. For this purpose, the WaterFlow Map was used again since the individual clusters can be actually interpreted as protein tunnels. The WaterFlow Map in Figure 9 helped to reveal the composition of individual clusters. It also allowed the biochemists to easily determine the most significant clusters and how they were used by the incoming and outgoing water molecules. Using the WaterFlow Map, the biggest cluster of interest was selected and the individual trajectories were displayed in 3D. Finally, the trajectory visualization was combined with DensityIsosurface and tunnels computed by the CAVER tool. The analysis was repeated for other clusters and combinations of clusters as well, providing a comprehensive overview of the structural protein tunnels with the water trajectories.

The case study presented here demonstrated the effectiveness of our proposed visualizations in analyzing the water trajectories and extracting the essential information out of the large amount of input data, i.e., water flows in and out of the active site. The navigation through our proposed TimeLine View and WaterFlow Map was fast and intuitive and significantly reduced the number of input water trajectories. Therefore, studying such a highly reduced set of trajectories in 3D became feasible. The DensityIsosurface representation provided an overview of the space occupied by selected water molecules over time. In combination with the information about the protein tunnels, the biochemists immediately revealed those tunnels, which were effectively used by the water molecules to access the active site. The WaterFlow Map also proved to be highly informative because it concisely summarized important information about the size of the clusters and how they were traversed by water molecules.

The biochemists also suggested that the proposed visualizations could be extended in the future and serve as a comparison tool for different molecular dynamics simulations. There are several known practical cases where protein engineers aim to change the protein properties and function by engineering transport tunnels. For example, Brezovský et al. [BBD*16] describe the computational design and directed evolution of a de novo transport tunnel in haloalkane dehalogenase. They presented mutants with blocked original tunnels and introduced newly-opened auxiliary tunnels. This change modified the protein properties dramatically. Molecular dynamics simulations confirmed the functionality of these auxiliary tunnels. When designing such mutations, it would be helpful to have the possibility to analyze the flow of water molecules. Here, the WaterFlow Map and the DensityIsosurface representation could provide the most valuable information on how the mutations might change the flux of water molecules, their clustered trajectories, and occupancy of water molecules inside the protein.

5. Conclusions and Future Work

In this paper we propose a set of visual representations serving for the visual analysis and inspection of large numbers of water trajectories in long molecular dynamics simulations. This is a very common situation in molecular biology studies. These representations guide the biochemistry researchers through the complex input data. They reveal the most interesting trajectories by studying their different features and behaviors. Based on the initial classification of trajectories, the user can interactively explore them using the TimeLine View. Different ranking possibilities reveal the most interesting trajectories, which can be further scrutinized using the WaterFlow Map. Here the set of trajectories is further divided into clusters according to protein entrance and exit points. The user can interact with the clusters and identify the correspondence between the incoming and outgoing parts of the trajectories. Next, the selected clusters can be visualized in a 3D view of individual trajectories (original or simplified ones) or as isosurfaces representing the occupancy of the protein inner region by the corresponding water molecules. All proposed visualizations are integrated into a visual analysis framework, where the user can interactively manipulate individual views and see the results in the other views. Currently one of the main bottlenecks, which we plan to focus on in the future, is the preprocessing stage where we have to analyze the positions of all water molecules in all time steps with respect to the protein.

Acknowledgements

This work was supported through grants from the Vienna Science and Technology Fund (WWTF) through project VRG11-010, the Czech Science Foundation (GACR) under research project No. 17-07690S, projects no. LO1214, LQ1605, and the PhysioIllustration research project 218023 funded by the Norwegian Research Council.

References

[AA11] Andrienko N., Andrienko G.: Spatial generalization and aggregation of massive movement data. *IEEE Trans Vis Comput Graph 17*, 2 (2011), 205–219. 2

[AAB*13] Andrienko G., Andrienko N., Bak P., Keim D., Wrobel S.: *Visual Analytics of Movement*. Springer, Heidelberg, 2013. 2

[ALC16] Alharbi N., Laramee R. S., Chavent M.: MolPathFinder: Interactive Multi-Dimensional Path Filtering of Molecular Dynamics Simulation Data. In *Computer Graphics and Visual Computing (CGVC)* (2016), pp. 9–16. 3

[BBD*16] Brezovský J., Babková P., Degtjarik O., Fořtová A., Góra A., Iermak I., Řezáčová P., Dvořák P., Smatanová I. K., Prokop Z., Chaloupková R., Damborský J.: Engineering a de novo transport tunnel. *ACS Catalysis 6*, 11 (2016), 7597–7610. 9

[BGB*08] Bidmon K., Grottel S., Bös F., Pleiss J., Ertl T.: Visual abstractions of solvent pathlines near protein cavities. *Comput Graph Forum 27*, 3 (2008), 935–942. 3

[BP14] Benson S. P., Pleiss J.: Solvent Flux Method (SFM): A Case Study of Water Access to Candida antarctica Lipase B. *J Chem Theory Comput 10*, 11 (2014), 5206–5214. 3

[CKS*16] Cornel D., Konev A., Sadransky B., Horvath Z., Brambilla A., Viola I., Waser J.: Composite flow maps. *Comput Graph Forum 35*, 3 (2016), 461–470. 2

[CPB*12] CHOVANCOVÁ E., PAVELKA A., BENEŠ P., STRNAD O., BREZOVSKÝ J., KOZLÍKOVÁ B., 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 Comput. Biol. 8*, 10 (2012), e1002708. 8

[CRG*14] CHAVENT M., REDDY T., GOOSE J., DAHL A. C. E., STONE J. E., JOBARD B., SANSOM M. S. P.: Methodologies for the analysis of instantaneous lipid diffusion in md simulations of large membrane systems. *Faraday Discuss. 169* (2014), 455–475. 3

[DV10] DEMŠAR U., VIRRANTAUS K.: Space-time density of trajectories: exploring spatio-temporal patterns in movement data. *Int. J. Geogr. Inf. Sci. 24*, 10 (2010), 1527–1542. 2

[DWL08] DODGE S., WEIBEL R., LAUTENSCHÜTZ A.-K.: Towards a taxonomy of movement patterns. *Information Visualization 7*, 3 (2008), 240–252. 2

[EKK*14] ERTL T., KRONE M., KESSELHEIM S., SCHARNOWSKI K., REINA G., HOLM C.: Visual analysis for space-time aggregation of biomolecular simulations. *Faraday Discuss. 169* (2014). 3

[FH75] FUKUNAGA K., HOSTETLER L.: The estimation of the gradient of a density function, with applications in pattern recognition. *IEEE Trans. Inf. Theory 21*, 1 (1975), 32–40. 4

[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), 22. 2

[Hol06] HOLTEN D.: Hierarchical edge bundles: Visualization of adjacency relations in hierarchical data. *IEEE Trans Vis Comput Graph 12*, 5 (2006), 741–748. 2

[KKF*16] 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. *Comput Graph Forum* (2016). 2

[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. *Comput Graph Forum 35*, 3 (2016), 527–551. 2

[KSES12] KRONE M., STONE J., ERTL T., SCHULTEN K.: Fast Visualization of Gaussian Density Surfaces for Molecular Dynamics and Particle System Trajectories. In *EuroVis - Short Papers* (2012), Meyer M., Weinkaufs T., (Eds.), The Eurographics Association. 4

[LC87] LORENSEN W. E., CLINE H. E.: Marching cubes: A high resolution 3D surface construction algorithm. *Computer 21*, 4 (1987), 163–169. 7

[LMS*12] LUBOSCHIK M., MAUS C., SCHULZ H.-J., SCHUMANN H., UHRMACHER A.: Heterogeneity-based guidance for exploring multiscale data in systems biology. *Proceedings of the IEEE Symposium on Biological Data Visualization (BioVis'12)* (2012). 3

[LRB*15] LUBOSCHIK M., RÖHLIG M., BITTIG A., ANDRIENKO N., SCHUMANN H., TOMINSKI C.: Feature-driven visual analytics of chaotic parameter-dependent movement. *Computer Graphics Forum 34*, 3 (2015), 421–430. 2

[LvB*17] LIŠKOVÁ V., ŠTĚPÁNKOVÁ V., BEDNÁŘ D., BREZOVSKÝ J., PROKOP Z., CHALOUPKOVÁ R., DAMBORSKÝ J.: Different structural origins of the enantioselectivity of haloalkane dehalogenases toward linear β-haloalkanes: Open-solvated versus occluded-desolvated active sites. *Angewandte Chemie 129*, 17 (2017). 1

[MGKH07] MATKOVIČ K., GRACANIN D., KONYHA Z., HAUSER H.: Color linesview: An approach to visualization of families of function graphs. *Information Visualization, 2007. IV'07. 11th International Conference* (2007), 59–64. 4

[MMG*16] MAGDZIARZ T., MITUSIŃSKA K., GOŁDOWSKA S., PŁUCIENNIK A., STOLARCZYK M., ŁUGOWSKA M., GÓRA A.: AQUA-DUCT version 1.0, http://www.aquaduct.pl/, 2016. 3

[MMG*17] MAGDZIARZ T., MITUSIŃSKA K., GOŁDOWSKA S., PŁUCIENNIK A., STOLARCZYK M., ŁUGOWSKA M., GÓRA A.: AQUA-DUCT a ligands tracking tool. *Accepted for publication in Bioinformatics* (2017). 3, 8

[PKC*09] PAVLOVÁ M., KLVAŇA M., CHALOUPKOVÁ R., BANÁŠ P., OTYEPKA M., WADE R., NAGATA Y., DAMBORSKÝ J.: Redesigning dehalogenase access tunnels as a strategy for degrading an anthropogenic substrate. *Nat. Chem. Biol.*, 5 (2009), 727–733. 1, 8, 9

[PXY*05] PHAN D., XIAO L., YEH R., HANRAHAN P., WINOGRAD T.: Flow map layout. In *Proceedings of the IEEE Information Visualization* (2005), pp. 219–224. 2

[RK11] RAUNEST M., KANDT C.: dxTuber: Detecting protein cavities, tunnels and clefts based on protein and solvent dynamics. *J. Mol. Graph. Model. 29*, 7 (2011), 895–905. 7

[Sch15] SCHRÖDINGER, LLC: The PyMOL molecular graphics system, version 1.8. November 2015. 3

[SG64] SAVITZKY A., GOLAY M. J.: Smoothing and differentiation of data by simplified least squares procedures. *Anal. Chem. 36*, 8 (1964), 1627–1639. 6

[SRE05] STEGMAIER S., RIST U., ERTL T.: Opening the can of worms: an exploration tool for vortical flows. In *Proceedings of IEEE Visualization* (2005), pp. 463–470. 2

[VCMB10] VASSILIEV S., COMTE P., MAHBOOB A., BRUCE D.: Tracking the flow of water through photosystem ii using molecular dynamics and streamline tracing. *Biochemistry 49*, 9 (2010), 1873–1881. 3

[VJN*15] VROTSOU K., JANETZKO H., NAVARRA C., FUCHS G., SPRETKE D., MANSMANN F., ANDRIENKO N., ANDRIENKO G.: SimpliFly: A methodology for simplification and thematic enhancement of trajectories. *IEEE Trans Vis Comput Graph 21*, 1 (2015), 107–121. 2

[VW90] VISVALINGAM M., WHYATT J. D.: The Douglas-Peucker algorithm for line simplification: Re-evaluation through visualization. *Comput Graph Forum 9*, 3 (1990), 213–225. 6