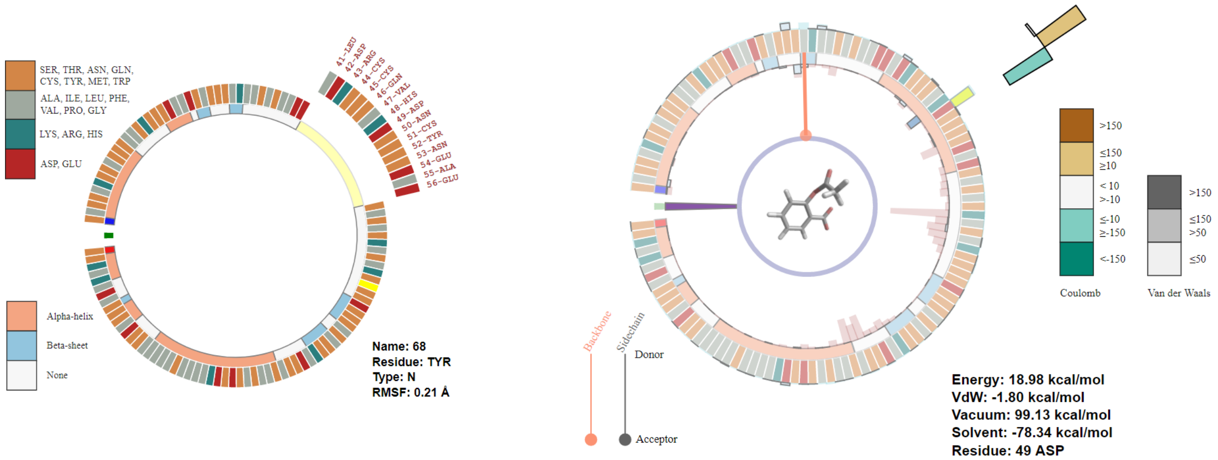
Eurographics Conference on Visualization (EuroVis) 2018
J. Heer, H. Leitte, and T. Ropinski
(Guest Editors)

Volume 37 (2018), Number 3

Visual Analysis of protein-ligand interactions

P. Vázquez[1], P. Hermosilla[2], V. Guallar[3,4], J. Estrada[3,5], and A. Vinacua[1]

[1] ViRVIG Group, Universitat Politècnica de Catalunya, Barcelona [2] Visual Computing Group, U. Ulm
[3] Barcelona SuperComputing Center [4] Institució Catalana de Recerca i Estudis Avançats (ICREA) [5] Bosonit, SL



Name: 68
Residue: TYR
Type: N
RMSF: 0.21 Å

Energy: 18.98 kcal/mol
VdW: -1.80 kcal/mol
Vacuum: 99.13 kcal/mol
Solvent: -78.34 kcal/mol
Residue: 49 ASP

Figure 1: *Two examples of our molecular interactions visualization. The left image shows the initial inspection of the secondary structures and the residues in the protein. By hovering on a secondary structure (top right) or residue (bottom right), we obtain the details. The right image shows the energy inspection mode, with the ligand shown in the center of the display. Here, the user can see both aggregate information, in the form of the maximum energies reached along the simulation path and inspect details about individual steps, such as the energy components (i. e. Van der Waals, Coulomb Solvent, and Coulomb Vacuum) of the interaction energy (by hovering onto the energy bars) or the intra-molecular bond that is formed (in orange) at this step of the simulation.*

Abstract

The analysis of protein-ligand interactions is complex because of the many factors at play. Most current methods for visual analysis provide this information in the form of simple 2D plots, which, besides being quite space hungry, often encode a low number of different properties. In this paper we present a system for compact 2D visualization of molecular simulations. It purposely omits most spatial information and presents physical information associated to single molecular components and their pairwise interactions through a set of 2D InfoVis tools with coordinated views, suitable interaction, and focus+context techniques to analyze large amounts of data. The system provides a wide range of motifs for elements such as protein secondary structures or hydrogen bond networks, and a set of tools for their interactive inspection, both for a single simulation and for comparing two different simulations. As a result, the analysis of protein-ligand interactions of Molecular Simulation trajectories is greatly facilitated.

CCS Concepts

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

1. Introduction

In many areas such as pharmacology or biotechnology, researchers use computers to perform molecular simulations (MS) of the interactions between molecules, as in, for instance, enzyme engineering and drug design. These simulations, involve different modeling techniques including docking, cavity detection, molecular dynamics (MD) [CFS14] and Monte Carlo (MC) [LW99] methods. In the context of drug design, MS aim at predicting the binding mode and affinity of a small molecule, the ligand, with a larger biomolecule, the protein. Such a binding may inhibit or activate certain function of the biomolecule (such as in the case of drugs), which results in a therapeutic benefit for the patient. In contrast with the large efforts carried out in the field of molecular structural analysis, with a number of software packages providing all sorts of geometric information on the molecules, much less effort has been devoted to the development of efficient tools for the visual analysis of other variables such as the interaction forces or the evolution of hydrogen bonds. Simulation outcomes are commonly examined by groups of scientists, in time consuming sessions where they typically discuss over a set of simple plots and (optionally) 3D renderings.

There is still a lack of tools providing visual information of the physico-chemical properties of Molecular Simulations. This is aggravated by the fact that due to the continuous decrease in computation costs and improvement in the computational power of present computers, simulation techniques generate larger and larger amounts of data. For instance, MS techniques easily generate hundreds of thousands of representative structures of a dynamical process. On top of that, richer information is stored per step. In this paper we address this problem with a set of visualization motifs designed specifically for the representation of the interaction variables of MS, and apply them to the concrete case of protein-ligand simulations. This visualization system allows to explore interactively a huge amount of physico-chemical information on molecular simulations. Moreover, the circular design provides a link to the 3D structure of the molecule by maintaining the spatial order of the secondary structures. Finally, the aggregated information lets the user grasp the essence of the simulation, and the interaction tools permit further exploration on the details of each residue in any particular simulation step.

As a result, researchers are able to quickly answer questions such as: **Q1**: Which are the most flexible residues of the protein?; **Q2**: How are protein-ligand interaction energies evolving through the simulation?; or more complex ones such as **Q3**: How different are these two simulations; **Q4**: Which residues are strongly interacting with the ligand at a certain distance?; or **Q5**: How do the intra(or inter)-molecular hydrogen bonds change along the ligand binding process? Some of these queries require the inspection of a certain plot (e.g. Q1), but others can only be carried out by having step-by-step information on certain simulation variables (e.g. Q2), being able to inspect several features (variables, trajectories) at once (e.g. Q3 and Q5), or having advanced filtering tools for the generated data (e.g. Q4). See subsection 6.2 for a more detailed comparison of our system and the limitations of current commercial software.

To sum up, our contributions are:

- A compact set of representations that simultaneously convey structural and functional information on protein-ligand interactions.
- A set of interaction tools and coordinated views to facilitate the inspection of whole MS trajectories.
- Multiple views that enable the comparison between different simulations (same protein and different ligand, or same ligand and different proteins).

To ensure uniformity and consistency, we base all of our inspection views on a circular representation of the protein sequence. This has two advantages: first, the circular representation facilitates the depiction of intra- and intermolecular hydrogen bonds in a compact way; second, it facilitates capturing the essence of the simulation at a glance (except for very large molecules).

The rest of the paper is organized as follows: next two sections analyze related work and introduce some fundamental biophysics background. section 4 presents our visualization system, with examples of use cases in section 5. In section 6 we report the feedback obtained from domain experts and discuss the advantages of our software with respect to commercial packages. Finally, we conclude in section 7 discussing ways to improve our application.

2. Related Work

Molecular visualization has been a field of intense work in the last decades. The interested reader can refer to the recent survey by Kozlikova et al. [KKF*17] for the state of the art in visualization of biomolecules and the one by Krone et al. [KKL*16] on the visual analysis of biomolecular cavities. With the evolution of high performance computing capabilities and GPU power, the size and complexity of representations has grown enormously. We are able to render now, with high quality shading, proteins of millions of atoms in realtime with a desktop PC [FKE13, GKM*15, LMAPV15]. In the following, we analyze works that add illustrative motifs for *molecular interactions*, systems devoted purposely on *protein-ligand interaction forces*, and other *related libraries*.

Visualization of molecular interactions. One of the areas that has received most attention is the visualization of tunnels and cavities, and the interactions of ligands within (e.g. [BJG*15, BLMG*16, FJB*17]). An example of a tool in this area is as CAVER Analyst that incorporates these visualization motifs [KSS*14]. Cipriano and Gleicher [CG07] illustrate charges over the molecular surface by stylizing both the surface shape and the charge values. Some software packages provide a means to overlap a set of semi-transparent spheres around atoms, used to color encode atom properties, e.g., Coulomb charges or hydrophobicity, such as in Vesta [MI08, MI11]. The Hyde software color codes atoms according to the total affinity energy [SHL*12], but the process takes several seconds. Günter et al. also focus on the atom level, where the signed electron density and reduced gradient fields are computed and then simplified to illustrate van der Waals and steric repulsion forces between atoms [GBCG*14]. More recently, Skanberg et al. have proposed to visualize energy interactions between atoms through diffuse interreflections computed for the surfaces of the atoms [SVGR16]. Falk et al. visualize molecule reactions by means of arrows augmenting paths representing molecule trajectories [FKRE09]. Grottel et al. on the other hand focus on the molecular surface and visualize electrostatic dipoles through color overlaid on the surface [GBM*12]. Khazanov and Carlson exploit tables to communicate molecule interactions [KC13]. Besides this molecule level visualization, they also communicate the interaction between ligand and binding site through modification of color and van der Waals radii on an atom scale. Furthermore, they also indirectly address the residue scale by performing this depiction individually for each amino acid. Sarikaya et al. also take into account the residue scale, by visualizing classifier performance with respect to protein chains on which a classifier has operated [SAMG14]. Finally, to communicate the differences of surface projected parameters, Scharnowski et al. propose to use deformable models [SKR*14]. Our goal is to provide a compact representation that conveys a larger set of data features (e.g. energy, RMSF values, h-bonds...), both covering single step and aggregate information. As a result, we require space efficient representations to avoid cluttering. Thus, we chose a set of coordinated 2D dynamically configurable views.

Communicating interaction energies. Most of the existing approaches that visualize such energies exploit 2D views. For example LigPlot+ generates 2D views of ligand-protein interactions for a static frame [LS11], whereby some idioms are shown in the 2D view to illustrate interaction forces. Similar results can be obtained by using LeView [Cab13], Maestro [Sch16] or PoseView [SMR06]. In all cases, the rendered views are limited to a single step of the simulation, and little or no interaction is possible, which in particular forbids further exploration. Furthermore, the projection techniques that are used make mental linking to the 3D structures difficult. LigandScout exploits several views to support creating screening databases [WL05]. It allows for the interactive creation of so called pharmacophores, which can be optimized for a certain ligand. LigandScout highlights the ligand's key features interacting with the protein, and supports surface coloring based on lipophilicity, HBD/A, or charge. The latter based on predefined scoring functions. The PLIP system is a web service that generates 3D projections for web browsers [SSH*15]. However, the interaction is limited to changing view parameters or the camera pose, and no filtering can be applied based on simulation results. Hermosilla et al. [HEG*17] depict 3D renderings of proteins interacting with ligands. The authors also provide energy visualizations, in the form of directed cones, and they support simulation paths. However, they do not provide any means to visualize aggregated information, nor other kind of features such as hydrogen bonds. Word et al. also provide 3D views of hydrogen atom contacts, allowing the user to determine the most suitable orientation of Glutamine and Asparagine side chains [WLRR99].

Commercial packages and other libraries. Well-known commercial packages such as Maestro [Sch16] or Avogadro [HCL*12] are provided as software packages to be downloaded. Nowadays however, there is a tendency to develop software that is available through the web. Notable mentions are Ligplot+ [LS11] or ProViz [JMS*16]. However, they have the limitation of allowing little or no interaction. Moreover, in some cases, the information of interacting forces is provided as a single sequential plot, since this has been the standard in other biological visualization tools (e.g. [MJD*16]). Similar to these tools, other libraries have been considered. For instance, our system is based on D3 and built using JavaScript, so it can be easily deployed and made publicly available. More related to our approach are libraries that have been developed recently, tailored for the representation of biological data (or even general data) in circular form, making use of JavaScript and D3, or Perl in their internals. Circos [KSB*09, ZMD13] is one example of those, and BioCircos [CCL*16], an evolution tailored to facilitate its use.

In terms of visualization techniques, and in contrast to the commercial software systems that typically show 1D depictions, or 3D images of a single simulation step, we provide compact 2D views that the user can configure (e.g. by changing the simulation step, or hovering over it to see detailed information on the energy components). And opposite to tools such as Circos, we provide an application, not a toolkit to create your own visualization.

3. Molecular Simulations Background

Molecular Simulations are a set of techniques that use computers to simulate whether a small molecule (ligand) can bind to a larger biomolecule (the protein) to activate or inhibit a certain biomolecule function. In drug design, ligands are developed to provide a therapeutic benefit for the patient, such as inhibiting the transmission of signals that communicate pain. The result of a MS is a trajectory with the information of the positions of the atoms of both compounds in each step, together with associated data such as the energy of the system (formed by both molecules and the water solvent) and the temporal interactions that are established between the protein and the ligand (hydrophobic, electrostatic, and intermolecular hydrogen bonds or h-bonds) and that guide the ligand towards the bound conformation, which they also help to stabilize.

Proteins are sequences of residues, where each residue can be one of a set of 20 possible aminoacids. All residues are divided into the atoms forming the backbone (that connects residues sequentially), and those forming the side chain, which gives specific physico-chemical properties to that residue that favor different types of interactions with the ligand. To understand the function of the protein, and the ligand binding, the 3D structure of the protein is key. This structure is based on a scaffold of α-helices and β-sheets (what is known as the secondary structure), kept in place by a network of internal h-bonds between backbone atoms. Binding may require small changes in this scaffold, so as to allow the entrance of the ligand to the binding site (the place in the protein where the bound conformation happens), or to create the specific disposition of atoms that define the binding site. Besides, the different elements of the secondary structure of a protein are typically used by domain experts as high-level reference points in the protein sequence. Some of the important aspects of MS are described next.

3.1. Simulation energy

The strength of the binding (also known as binding affinity) is determined by the free energy of binding, which can be approximated as the sum of the interaction energies between the protein residues and ligand. Conformations with low energy levels (negative, with high absolute values) are more stable, and the real bound conformation will probably lie within one of those low energy configurations. The interaction energy of each residue with the ligand is additive and can be divided into van der Waals energy (a short-range interaction), electrostatic energy in the vacuum (long-range) and the screening of that electrostatic energy due to the solvent. Different residue types will favor different interactions and, therefore, binding to different ligands. For example, positively charged residues will favor interaction with negatively charged ligands. Our application currently deals with three components of the binding energy: *i)* Van der Waals energy (E_{VDW}), *ii)* electrostatic interaction energy in the vacuum (E_{vac}, also referred to as Coulomb vacuum), and *iii)* solvent screening of the electrostatic interaction energy (ΔG_{solv}, also called Coulomb Solvent). These terms are added up to compute the total binding energy:

$$\Delta G_{bind} \approx E_{VDW} + E_{vac} + \Delta G_{solv} \qquad (1)$$

In our implementation, E_{VDW} and E_{vac} are computed using the OPLS force field [BBC*05], while ΔG_{solv} is calculated using the generalized Born model [BC00]. Note that each MS software uses a different version of the binding force fields. We have chosen this energy breakdown, to illustrate the capabilities of the visualization, but more sophisticated energy terms might be shown as well.

3.2. Root Mean Square Fluctuation

The forces derived from the (potential) energy induce movements both in the protein and the ligand, these are commonly measured using the Root Mean Square Fluctuation (or RMSF). This quantity measures the average fluctuation of a residue's atom around a reference position. Residues with high values of this quantity may indicate that they move to interact with the ligand guiding it towards the binding site, to make room for the ligand, or to establish stabilizing interactions with the ligand. RMSF is used in different areas, notably as a quantitative measure of similarity between protein structures, but also to analyze which parts of a protein are affected by the interaction with ligands in MS. Its units are Angströms, and it is commonly measured as (using the notation by Maestro):

$$RMSF_i = \sqrt{\frac{1}{T}\sum_{t=1}^{T}\left(r_i'(t) - r_i(t_{ref})\right)^2} \qquad (2)$$

3.3. Hydrogen bonds

Inter-molecular hydrogen bonds, also referred to in literature as h-bonds, play a significant role in protein stability and ligand binding. Hydrogen bonds are formed between a hydrogen atom (the donor part of the bond) and a non-hydrogen atom (the acceptor), and are detected following geometric rules. When characterizing binding, the domain experts want to know whether the donor and acceptor in the protein is an atom belonging to the backbone or to the side-chain; therefore, we classify h-bonds in backbone acceptor, backbone donor, side-chain acceptor, and side-chain donor. In drug design, these are relevant because of their strong influence on drug specificity, metabolization, and absorption. For example, in the case of specificity, only ligands with a specific pattern of h-bonds with the protein in the bound conformation will have a strong binding energy. A domain expert may design new drugs by copying that h-bond pattern, or may understand that mutations of proteins that change residues involved in that h-bond pattern affect the binding rendering the drug ineffective. H-bonds are also important in determining the secondary structure of proteins, as explained before.

Intra-molecular hydrogen bonds are other, important h-bonds within the protein molecule itself. These play an important role in determining the three-dimensional structure adopted by proteins, because these bonds cause the molecule to fold into a specific shape. Actually, secondary structures are determined by such bonds: regularly occurring bonds between aminoacids at relative positions of i and $i+4$ form an alpha helix. On the contrary, beta sheets are formed when aminoacids of different strands are involved in hydrogen bonds.

4. Visualization setup

In this section we present the different graphical components of our system and how they are used to inspect molecules. The system consists of five components: *i)* secondary structures, *ii)* bonds, *iii)* energies, *iv)* RMSF, and *v)* energy chart. The first four share a circular shape that represents the chain of amino-acids that form the protein, and the fifth is a chart. These graphical representations may show information of a **single simulation step**, or add **aggregated information of a full trajectory**, which is very useful for getting the essence of full simulations. As described in subsection 4.5, our system shows several of those views, together with the energy chart, and provides tools for data filtering, as well as showing details on demand.

In the following, we describe the graphical components, and then the whole system. For compactness and improve legibility, in images depicting the different views, we moved legends closer to the visual depictions, so that we can enlarge the whole image.

4.1. Secondary structures

The first view conveys the structural properties of the protein, by showing its backbone (the open chain of its aminoacids) in an arc. The residues that belong to each individual secondary structure units are represented as bars in a second, external ring. This is shown in the left part of Figure 1. The color encoding of the secondary structures identifies them as either belonging to the α-helix (in orange) or the β-sheet (in light blue). There is no unified color coding in literature, but red-orange and blue-green are quite common. The dark red and dark blue indicate the beginning and ending of the backbone, respectively. Hovering over the different regions of the backbone brings up information on the residues involved (top left), and hovering over a single residue (bottom right) provides details of that residue. Note that the example also contains a calcium ion (in green on the left) that does not belong to the backbone. Thus, co-factors other than the ligand can be displayed as separate residues. The colors of the residues have been selected so that they communicate their standard polarity; residues in reddish tones are negatively charged, while those in blueish tones are positively charged, and the grey ones are residues with high hydrophobicity. The rest are in brown, representing their higher potential for hydrogen bonding.

4.2. Hydrogen bonds

For h-bonds visualization, we de-emphasize secondary structures and residues, add the ligand in the center of the representation, and overlay the h-bonds as arcs that point to the concrete residue interacting with the ligand. The circle indicates which molecule (ligand

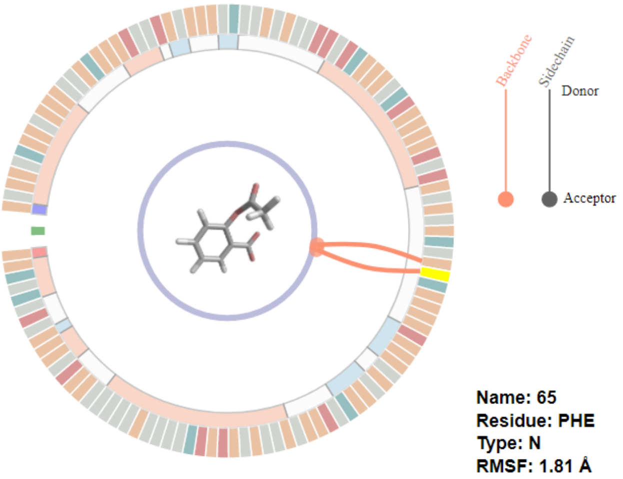


Figure 2: *Hydrogen bonds visualization. The dynamically created bonds between the drug and the protein are shown using a line and a circle. The position of the circle (next to the ligand or to the backbone) encodes the acceptor, and the color of the linking encodes the type of bond, side chain (grey) or backbone (orange).*

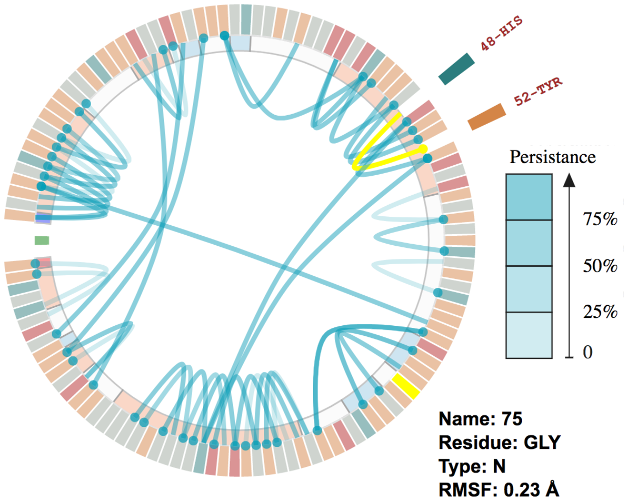


Figure 3: *Inter-molecular hydrogen bonds. The transparency level encode the bonds persistence, and hovering over a bond (top-right) or a residue (bottom-right) reveals its details.*

or protein) is the acceptor. The color indicates whether the connection is with the backbone (orange) or the side chain (grey), see Figure 2. Furthermore, h-bonds may be multiple. In such case, we multiply the number of links accordingly. It is important to take into account that, like energies, these bonds are dynamic, so they change along the path. The user can freely select a certain step of the path, or re-play the simulation continuously.

Intra-molecular bonds are represented similarly, but we change the color to make them more easily distinguishable, see Figure 3. Their regular structure (connecting aminoacids at positions i and $i+4$) can be easily identified. The existence of such bonds is also dynamic, and it is important to communicate their persistence along the simulation, thus, we encode this aggregated information with a higher degree of opacity of the bonds. The persistence is here measured as the percentage of the steps of the simulation in which the hydrogen bond was present. Details are obtained hovering over the bonds, which facilitates identifying the residues that intervene (top-right) as well as to recognize the donor and the acceptor (the one with the circle). Hovering over the residue (bottom-right) provides more details on the residue itself.

4.3. Energies

The energies are encoded in the following way: bars pointing to the center represent attracting forces, while bars outside the residues arc, encode repulsive forces. Their color encodes the magnitude, following a ColorBrewer palette [HB03]. Finally, in the background, if selected by the user, we also provide aggregate information of the extremes of the energy reached along the whole trajectory, see Figure 1-right. By hovering over each energy bar, the user gets detailed information on the total interaction energy, and the values of the components (top-right). These energy components (Van der Waals and electrostatic factors) for each residue, like the total value, point to or away from the ligand depending on their attractive or repulsive nature, respectively. These components are color-coded (see Figure 5 bottom-left) following a diverging palette also extracted from ColorBrewer's system. The detail legends also appear only on hovering, while displaying the energy disaggregation. The user may also click on a bar representing an extremum of the energy to jump to a step in the simulation where that extremum value is achieved.

4.4. Root Mean Square Fluctuation

Instead of visualizing this information as charts (as happens in common applications, e.g. Maestro), we render it next to the protein backbone, as shown in Figure 4, to facilitate relating the fluctuations with the residues involved. This way, we can provide much more information than in a chart, since we can see the information on the residues, the secondary structures they belong to and so on. As shown later, another advantage of our system is that it allows for multiple overlays over the same space, and cross-view highlighting. This permits easily and intuitively establishing visual relationships between different factors.

4.5. Overview of the system

For MS analysis, our system has three different setups: single protein-ligand simulation, same protein and different ligands, or two different proteins. All of them use the same components, though the last two require two energy charts instead of one. We analyze in detail the first one, depicted in Figure 5. We follow Schneiderman's mantra of overview first, zoom and filter, details on demand.

Initially, the system starts with the information on the protein (secondary structures) and the RMSF view, together with the energy chart. This gives an overview of the simulation, and provides

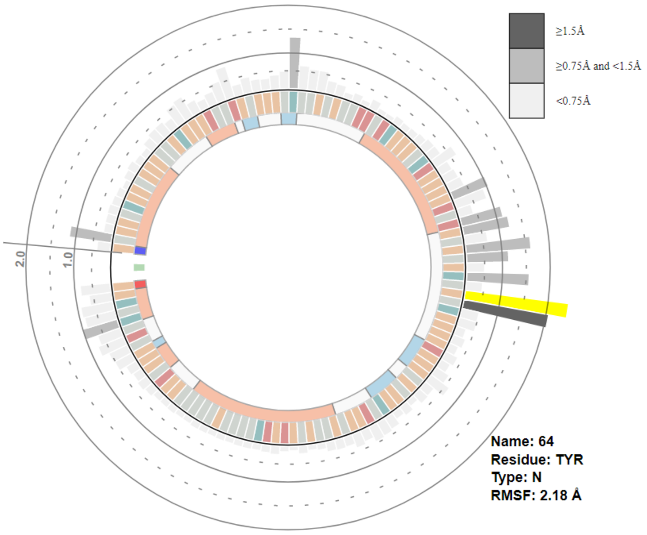


Figure 4: *Root Mean Square Fluctuation (RMSF) encoding. This component shows the aggregated information of the RMSF along the path, overlaid to the molecule structure. This is, in turn, de-emphasized to avoid catching the user's attention. By hovering over a bar, the details are revealed (bottom right).*

tips for further exploration. The user can change the view type with the drop-down menus, filter the data using the filtering tools (top-center), and get details by hovering on different parts of the visualization (e.g. in the central part of the image, details of the energy components are provided on demand). Moreover, both views are coordinated, so selection on one view triggers selection on the other. The same happens with the chart, that can be used as input widget, by dragging over it, the selected step is changed and so is done in the other views coordinately. The accompanying video shows the interaction with the application in more detail.

4.6. Multiple steps analysis

From the analysis of the initial stages of the application, the experts in our team suggested the addition of aggregated information in single views. Thus, we created several views that facilitate viewing the outcome of a whole simulation: the RMFS view, the overlaid min-max energies, and the intra h-bonds persistence. These facilitate solving some of the questions referred to as interesting for the experts. For example, with the RMSF view, we can solve the **use case** mentioned in **Q1**: learning which are the most mobile parts in the protein. In Figure 4, the RMSF bars clearly show which residues need to change position in order to allow the ligand to access the binding site (note the high values of the Tyrosine (TYR), highlighted, and the Phenylalanine (PHE) amino-acid, just below it). More complex questions can be solved by combining this view with others, as shown later in section 5.

The application allows mapping the individual intermolecular energy interactions between a ligand and all the residues, but also overlaying the maximum values reached for each residue. This can be used as reference, since it gives an idea of the potential for a specific interaction, comparing it with that of the current step (protein-ligand conformation). These values, for example, may be exploited by changing the structure of the ligand or by mutating the protein. They may also provide a hint as to which are the likely anchorage points for the ligand to the binding site, since those would probably get a value close to the maximum and stay near that value before other interactions are completely formed, as shown in section 5.

4.7. Multiple Coordinated Views

The use of multiple views increases the possibilities of the visual analysis, placing the different pieces of information together, helping the user establish connections between the step per-residue interaction energies, the energy components for selected residues, the maximum interaction energies per-residue during the whole simulation, and so on. Our basic application shows at least 3 coordinated views: two showing different aspects of the molecular simulation, and a third one showing the energy chart. Thanks to the modular design, we have also straightforwardly implemented two additional versions that extend the possibilities beyond the analysis of a single simulation. The first one inspects two simulations of the same molecule with different ligands. The second deals with two different proteins at once, with the same or different ligand. In those cases, instead of a single chart, we display two energy charts, with the same features as the previous one: the user can play each of the paths with the buttons or hovering over the corresponding chart. Clicking over a point in a chart also jumps to the corresponding step of the associated MS.

We provide several interaction ways between views especially designed to highlight correlated information. When the views show the same protein, clicking on one element in one view, highlights the corresponding information of the same residue in the other view, facilitating thus the analysis of the simulation information. Moreover, in those cases, the path is played in parallel in both views. When the proteins are different, the coordination happens between the main view and its corresponding chart, though other cross-view highlights could be incorporated if found necessary.

Not only single trajectory exploration is useful for researchers, but also comparing different trajectories or even different proteins, can also be of great utility. In the case of *the same protein-ligand system*, they can analyze which are the significant differential interactions between two bound poses for the ligand. In enzyme engineering, it is also typical to study *the same enzyme with different ligands* (called substrates), to understand why affinity for one substrate may be higher than for the other. It is also of interest to understand why protein mutation (for example, in virus proteins) can influence the effectiveness of a drug. And even *for different proteins and ligands*, the user may be querying whether a common interaction scheme underlies the binding in both protein-ligand systems. In the next section we show some examples of use cases.

5. Use cases

In this section we discuss how our system enables the analysis of several examples of interest, and show how it helps answering questions like those posed in section 1. In the analysis of these examples, the use of multiple views, as well as overlaid information,

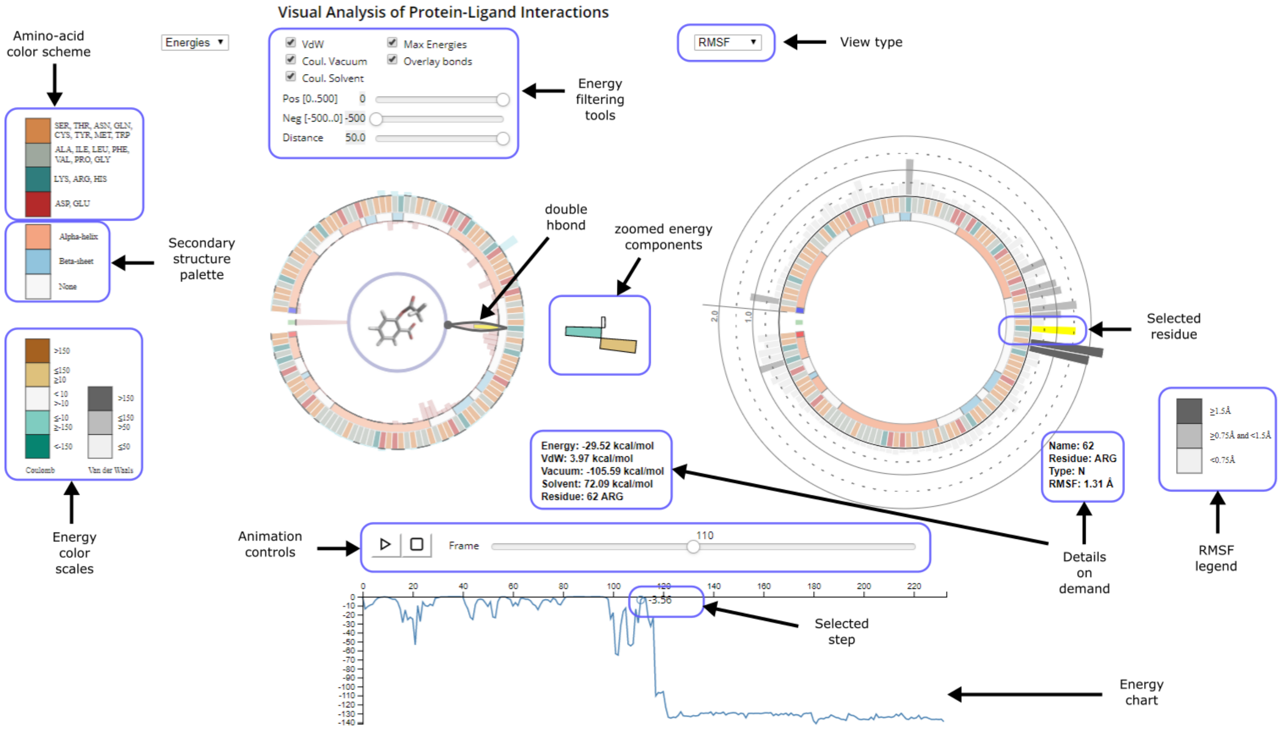


Figure 5: *A global view of the application with most of the features toggled on. The left half shows the main view, where the user will spend most of her time analyzing a molecule. Centered below is the slider controlling the steps of the MS. During the visual inspection process, the secondary view (right) can be used for comparing different views of the same molecule, or different molecules, as explained below. The bottom part shows the energy chart, which can also be used to control the step under scrutiny by moving over the curve.*

and cross-highlighting crucially enable the quick detection of key features of the simulations. Experts asked to evaluate our proposal (see subsection 6.1) found the multiple view setup highly useful. For completeness, the cases used involve simulations on the PDB protein-ligand models 1OXR (for the phospholipase-aspirin binding; [SEJ*05]), 1TD7 (for the phospholipase-niflumic acid binding; [JSS*05]), and 1A28 (for the progesterone nuclear hormone receptor-progesterone binding; [TWWS98]).

5.1. Analyzing energy view + energy plot

One of the elements we added to our system is the total interaction energy plot. It is the classical way to analyze an MS, and it can be used in our application as an interaction widget. By making the plot act as an input element, the user may quickly reach interesting parts of the simulation. In our case, hovering over the chart automatically sets the current step. This, combined with aggregated information, such as the min-max energies, facilitates understanding how a certain simulation has fared. We can see an example in Figure 6, where phospholipase-niflumic acid binding is shown. The combination of the total interaction energy of the system as shown in the chart with the information on the maximum and minimum reached energies by each residue, added to the instant energy analysis facilitates the exploration of the simulation outcome.

In this trajectory, as seen in the binding energy vs. step graph below the interaction energy plot, there are several minima found during the simulation; however, none of them corresponds to the much lower energies of the bound conformation (found in a different trajectory not shown here). This could be analyzed solely with the plot. However, in our case, we may further examine the detail of the energy simulation by residue, by using the energy view. As a result, we facilitate the understanding on the details of a certain simulation, as proposed in question **Q2** enunciated in section 1.

The energy view filtering tools can also be used to find answers to the **use case** referred to as **Q4** in section 1: finding highly interacting residues at a certain minimum distance. The filtering widgets can be used to define distance ranges. Then, with the min-max energy on, by clicking on the maximum/minimum energy bar, the step quickly changes to the desired step. Further analysis, either by moving back and forth the trajectory, and the extra information provided by the other view (e.g. bonds) helps to better understand the outcome of the simulation.

5.2. Combining h-bonds with RMSF

Besides the individual views, we can also add overlaid information over the views that facilitate the in-situ analysis of multiple variables. The motifs have been carefully designed so that in several cases, overlaying and blending produce readable depictions.

Since protein-ligand interaction may induce changes in the pro-

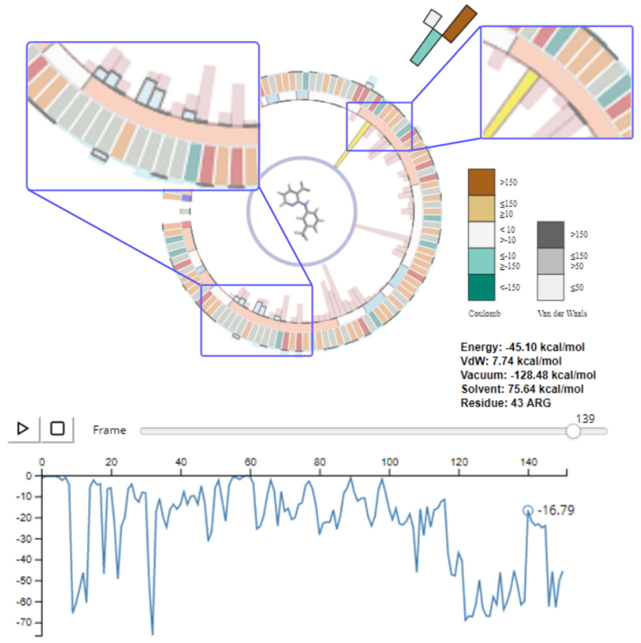


Figure 6: *Analysis of a binding study of phospholipase with niflumic acid. The top image shows a stage of the simulation close to the binding position. It is simple to detect that just a few residues are involved in the binding process and with low energies (see the insets), which may indicate a loose binding. Scrolling through the simulation, one can see that the attractive energies come from different residues along the path, thus not generating a strong binding. This can also be seen in the energy chart on the bottom, where the energy reduces, but not as much as in other cases analyzed.*

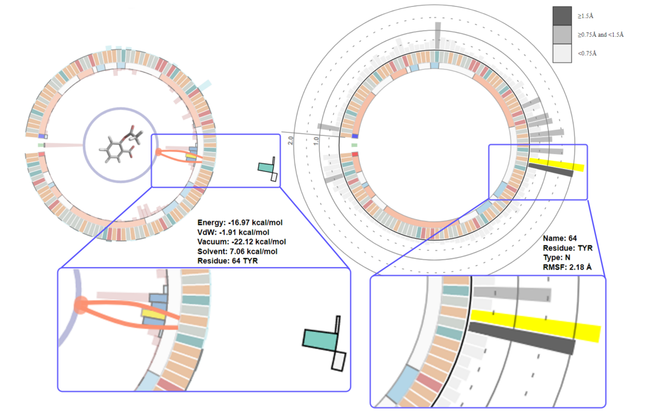


Figure 7: *The RMSF view and the final energies view found by interacting with the RMSF and min-max energies. The left view shows that, by searching for mobile parts on the RMSF view, we found a double step h-bond that is formed due to the attractive force exerted by the Tyr64 amino acid.*

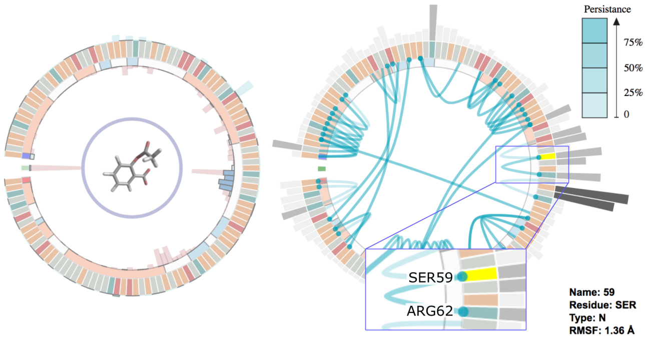


Figure 8: *Capturing the protein-ligand recognition motif of the aspirin to the phospholipase protein through the analysis of intramolecular bonds. -The serine residue (selected) does not form a persistent bond that sometimes creates a secondary or a tertiary structure, but an h-bond with low persistence that may have formed to favor the binding. The RMSF around the Arg62, bound to it, indicates high fluctuation, suggesting the idea of an induced fit between the protein and the ligand in that region.*

tein shape, one may be interested in analyzing whether a highly mobile part of a protein may be part of the ligand-recognition mechanism, driving the ligand closer to the active site. This can be solved using the RMSF and energy. In this last case, we configure the energy view to show the min-max energies, overlaid with the h-bonds. The procedure is simple: By clicking on the high mobile part, indicated by the longest RMSF bar on the right of the protein, the user can identify the corresponding residue as the tyrosine amino acid (Tyr64) on the left view. Then, by clicking on the minimum energy, the system automatically jumps to the corresponding step of minimum (most attractive) energy. And it results that in this step a double h-bond is formed, due to the attractive force of Tyr64 and its neighbor Phe65. Figure 7 shows the solution obtained in the energies view of this case, and the RMSF view used to find it (right).

Multiple views analysis helps to solve also the **use case** mentioned as **Q5**: the analysis on how the h-bonds evolve along the binding process. The procedure is as follows: the RMSF views lets us see other residues that stand out due to their high fluctuation. By hovering over the RMSF bars we see that in this case it is glycine Gly30. When clicking on the energies chart for the minimum energy, we see that in this step it does not form any hydrogen bond to the ligand. Then, scrolling the hydrogen bond along the protein-ligand dynamics (using the energy chart), we observe that a hydrogen bond finally forms, in what was identified as the bound conformation, inferring an additional stabilizing role. This is consistent to the binding procedure as noticed in [SEJ*05], where it is found that this residue forms part of the binding site for the calcium ion, and may move to make room for the binding ligand (and thus its high mobility).

5.3. Analyzing the intra h-bonds + RMSF

To find the recognition motif facilitating the binding of the ligand to the protein, the user can choose to visualize (as in Figure 8) a composite of the intramolecular energies (on the left of the image) and the intramolecular bonds, with the RMSF overlaid (right-hand side). By selecting the large electrostatic component of the arginine in the left view (purple bar on the inside of the wheel, just below the equator, on the right hand of the wheel), the user is taken to the frame that exhibits that energy, where the right-hand wheel shows the formation of a low-persistence bond between the arginine and the serine residue (selected, and consequently highlighted in yellow); this suggests that the bond may have formed to favor the binding. The gray bars on the outside of the right-hand wheel show that both this arginine and its neighbors present a relatively high fluctuation, further supporting the idea of an induced fit between the protein and the ligand affecting that protein region.

5.4. Protein comparison

The software also allows to set up side-by-side views of different interactions to compare them. In Figure 9 we can see a comparison of two different binding processes. The energy plot of the left hand simulation (binding of aspirin to Phospholipase A2) shows a sharp drop in the energy of the system just before step 120; after that, the energy stabilizes at a strong negative value, suggesting a bound state (as really is the case). Hovering over the energy chart, the user can inspect the evolution of the interaction energy components. In the figure, the user has stopped over a point near the formation of this configuration, and can see at the wheel a large electrostatic interaction with the calcium ion. This represents a classical example of a charged ligand and protein (a calcium cofactor here), where electrostatic complementarity accounts for most of the bounds interaction energy.

A different situation can be seen at the right-hand side display (progesterone nuclear hormone receptor with progesterone), where the energy plot shows that the binding interaction is the sum of several small interactions. Hovering over each of them, we can see their different nature, and the magnitude of the Van der Waals and electrostatic terms. In the figure we have selected one of the final steps of a binding trajectory, corresponding to a conformation close to the bound state as seen from an X-ray experiment. This shows completely different interaction profile, with several residues having interaction energies close to their maxima, and with several distant regions (in sequence) involved in the binding.

6. Evaluation

6.1. Feedback from domain experts

Throughout all the development of the project, we used input from two domain experts to set up the requirements for solving their problems. This input caused an increase in the number of features and helped improve the overall design of the application. However, we also gathered the opinions of three other experts outside the group of collaborators. We have provided our application to these external domain experts, and asked them to use it for 20 to 30 minutes, with a simple manual briefly describing the main goal and the features of the application. Then, they were asked to answer a questionnaire about the utility of the application and their interest in using such software. In all cases, the experts **are willing to use such an application in their work**, they were also unanimous in assessing that **such a tool helps them to better understand the overall behavior of a simulation**, and to infer h-bonds formation. They also found especially useful **to summarize trajectories**, and thought that this **is not available in this form in known software** ("like VMD, Maestro, MOE, Chimera, Pymol, Chem3D, etc."). As extra features, they would add some more properties and a 3D view coupled with it. Given that our software is developed in D3, there are ways to achieve this. First, a Qt application can embed the D3 code and render the 3D view in a new window (in Figure 10, for instance, we show the corresponding step of the simulation in a Qt window). However, this may reduce the possibilities of making the application widely accessible or used as SaaS. A second way would be to move our 3D rendering code to WebGL and embed the view into the application.

6.2. Comparison with other software

The system presented is able to answer a wide range of questions with little effort. Some of them can be analyzed by current commercial software, such as Maestro's SID or VMD's Timeline plugin, for **Q1** or **Q3**. Other problems such as **Q2** (analyzing the protein-ligand interactions along the simulation) are more difficult to answer. In this case, current packages such as VMD or LigPlot+, will yield the total energy, but not break them down into specific terms. Others, such as Maestro, will provide the energy terms for a single step, not allowing to move back and forth the simulation. The same happens with **Q4** (the examination of residues interacting with a ligand at a certain distance). Commercial software will restrict the analysis of interactions at close distances (e.g. below 3 Angströms), thus, do not allow filtering by distances. The analysis of the evolution of intra or inter-molecular h-bonds (example **Q5**) can only be done with commercial software in 3D. Occlusions in 3D make very difficult for researchers to grasp the information. 2D versions, such as the one included in Maestro, do not allow the exploration of a whole trajectory, just a single step. More in general, VMD lacks the side-by-side comparison of multiple variables, and Maestro's SID does not provide trajectory comparison on integrated view of the key features involved in binding, since the information is separated into different 2D plots.

7. Conclusions and Future Work

We have developed a new system for the visual analysis of molecular structures and interactions using D3, with the input data stored as XML files, no further requirements in terms of hardware or software are needed. We remove most of the physical information of the molecule, and leave just the protein sequence as basis, rendered in an arc, along the residues that belong to each part of the backbone. Then, we use InfoVis techniques to provide a high amount of information on the simulation, and several ways to visually explore and analyze the trajectory. However, the element that makes our application different from others is the possibility of analyzing and exploring multiple views at once. With several aggregated views including minimum and maximum energies, Root Mean Square Fluctuation, or the persistence of hydrogen bonds along a simulation path, the user can easily grasp the outcome of a simulation, and gather important details. For example, the minimum and maximum view allows the observer to determine which sets of residues have been active (or inactive) throughout a simulation, as well as the ones which have exerted more important forces.

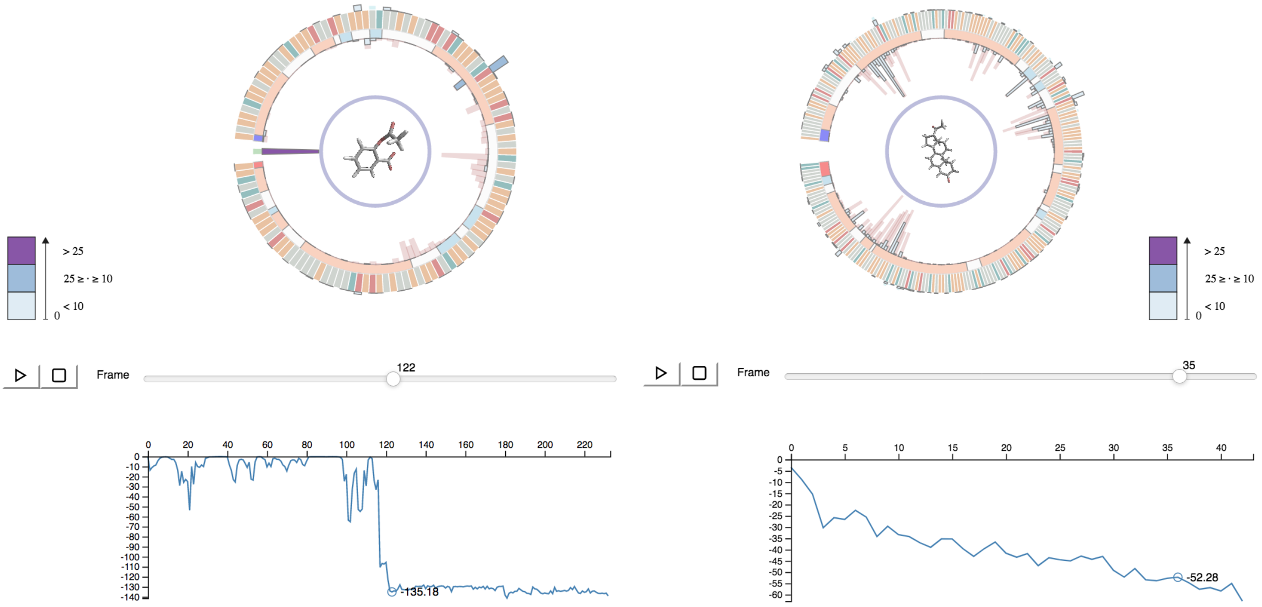


Figure 9: *Comparison of two different binding studies: binding of aspirin to Phospholipase A2 (left), and progesterone nuclear hormone receptor and progesterone (right). With a simple analysis, one can detect that in the first case, just a few residues are involved in the binding process, but one of them exhibits a very strong attractive energy. The progesterone nuclear hormone receptor behaves completely opposed to this, with many residues acting with levels of attraction close to their maximum, suggesting a good binding.*

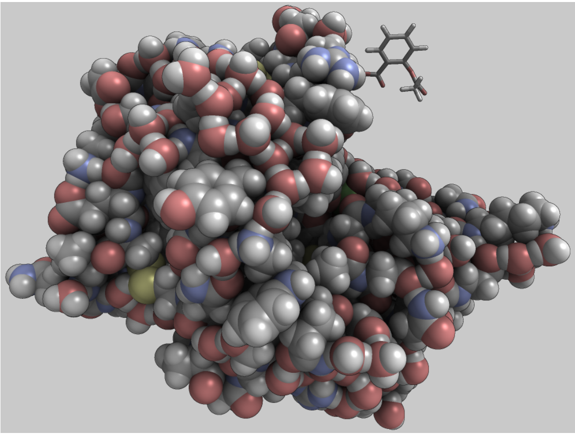


Figure 10: *3D view of a concrete frame viewed using a Qt window.*

In addition, we have added the possibility of inspecting multiple views at once (actually, the modular design of our application would let us add more coordinated views, provided we had enough screen space). This brings the possibility of analyzing, not only different variables of a single simulation, but to compare two different simulations, either with the same protein-ligand pair, comparing the outcomes of two simulations with the same protein and different ligands, (shown briefly in the accompanying video), or even analyzing two completely different molecules and drugs (see the accompanying video). The flexibility we provide is currently not available in other related software. However, there is still more information that can be added to our visual analysis tool, such as ligand RMSD, distance to the ligand... In the future we want to add other information such as contacts or torsion profiles, as well as more fine grain in the detail inspection, for instance, identifying the concrete atom in a certain bond or having per-frame information of the fluctuation. The latter is simple to add as soon as we have the data, the former will require some design changes to make room for the detailed information.

Acknowledgments

The authors want to thank the anonymous reviewers for their comments that greatly helped improve the paper. This project has been supported by the *Spanish Ministerio de Economía y Competitividad* and by 839 FEDER (EU) funds under the Grants No. TIN2014-52211-C2-1-R and TIN2017-88515-C2-1-R (GEN3DLIVE), and CTQ2016-79138-R (V-Mutate). This work was also partially funded by the Deutsche Forschungsgemeinschaft (DFG) under grants RO 3408/2-1 (ProLInt) and RO 3408/3-1 (Inviwo).

References

[BBC*05] Banks J. L., Beard H. S., Cao Y., Cho A. E., Damm W., Farid R., Felts A. K., Halgren T. A., Mainz D. T., Maple J. R., et al.: Integrated modeling program, applied chemical theory (impact). *Journal of computational chemistry 26*, 16 (2005), 1752–1780. 4

[BC00] Bashford D., Case D. A.: Generalized born models of macromolecular solvation effects. *Annual review of physical chemistry 51*, 1 (2000), 129–152. 4

[BJG*15] Byška J., Jurčík A., Gröller M. E., Viola I., Kozlíková B.: Molecollar and tunnel heat map visualizations for conveying spatio-temporo-chemical properties across and along protein voids. In *Computer Graphics Forum* (2015), vol. 34, Wiley Online Library, pp. 1–10. 2

[BLMG*16] Byška J., Le Muzic M., Gröller M. E., Viola I., Kozlikova B.: Animoaminominer: Exploration of protein tunnels and their properties in molecular dynamics. *IEEE transactions on visualization and computer graphics 22*, 1 (2016), 747–756. 2

[Cab13] Caboche S.: Leview: automatic and interactive generation of 2d diagrams for biomacromolecule/ligand interactions. *Journal of cheminformatics 5* (2013), 40. 3

[CCL*16] Cui Y., Chen X., Luo H., Fan Z., Luo J., He S., Yue H., Zhang P., Chen R.: Biocircos. js: an interactive circos javascript library for biological data visualization on web applications. *Bioinformatics* (2016), btw041. 3

[CFS14] Ciccotti G., Ferrario M., Schuette C.: Molecular dynamics simulation. *Entropy 16* (2014), 233. 1

[CG07] Cipriano G., Gleicher M.: Molecular surface abstraction. *IEEE Transactions on Visualization and Computer Graphics 13*, 6 (Nov. 2007), 1608–1615. URL: http://dx.doi.org/10.1109/TVCG.2007.70578, doi:10.1109/TVCG.2007.70578. 2

[FJB*17] Furmanová K., Jarešová M., Byška J., Jurčík A., Parulek J., Hauser H., Kozlíková B.: Interactive exploration of ligand transportation through protein tunnels. *BMC bioinformatics 18*, 2 (2017), 22. 2

[FKE13] Falk M., Krone M., Ertl T.: Atomistic visualization of mesoscopic whole-cell simulations using ray-casted instancing. *Computer Graphics Forum* (2013), 195–206. doi:10.1111/cgf.12197. 2

[FKRE09] Falk M., Klann M., Reuss M., Ertl T.: Visualization of signal transduction processes in the crowded environment of the cell. In *Proceedings of the 2009 IEEE Pacific Visualization Symposium* (Washington, DC, USA, 2009), IEEE Computer Society, pp. 169–176. URL: http://dx.doi.org/10.1109/PACIFICVIS.2009.4906853, doi:10.1109/PACIFICVIS.2009.4906853. 2

[GBCG*14] Gunther D., Boto R. A., Contreras-Garcia J., Piquemal J.-P., Tierny J.: Characterizing molecular interactions in chemical systems. *Visualization and Computer Graphics, IEEE Transactions on 20*, 12 (2014), 2476–2485. 2

[GBM*12] Grottel S., Beck P., Muller C., Reina G., Roth J., Trebin H.-R., Ertl T.: Visualization of electrostatic dipoles in molecular dynamics of metal oxides. *IEEE Transactions on Visualization and Computer Graphics 18*, 12 (Dec. 2012), 2061–2068. URL: http://dx.doi.org/10.1109/TVCG.2012.282, doi:10.1109/TVCG.2012.282. 2

[GKM*15] Grottel S., Krone M., Muller C., Reina G., Ertl T.: Megamol – a prototyping framework for particle-based visualization. *Visualization and Computer Graphics, IEEE Transactions on 21*, 2 (Feb 2015), 201–214. 2

[HB03] Harrower M., Brewer C. A.: Colorbrewer. org: an online tool for selecting colour schemes for maps. *The Cartographic Journal 40*, 1 (2003), 27–37. 5

[HCL*12] Hanwell M. D., Curtis D. E., Lonie D. C., Vandermeersch T., Zurek E., Hutchison G. R.: Avogadro: An advanced semantic chemical editor, visualization, and analysis platform. *J. Cheminformatics 4* (2012), 17. 3

[HEG*17] Hermosilla P., Estrada J., Guallar V., Ropinski T., Vinacua A., Vázquez P.-P.: Physics-based visual characterization of molecular interaction forces. *IEEE transactions on visualization and computer graphics 23*, 1 (2017), 731–740. 3

[JMS*16] Jehl P., Manguy J., Shields D. C., Higgins D. G., Davey N. E.: Proviz—uta web-based visualization tool to investigate the functional and evolutionary features of protein sequencesu. *Nucleic acids research 44*, W1 (2016), W11–W15. 3

[JSS*05] Jabeen T., Singh N., Singh R. K., Sharma S., Somvanshi R. K., Dey S., Singh T. P.: Non-steroidal anti-inflammatory drugs as potent inhibitors of phospholipase a2: structure of the complex of phospholipase a2 with niflumic acid at 2.5 å resolution. *Acta Crystallographica Section D: Biological Crystallography 61*, 12 (2005), 1579–1586. 7

[KC13] Khazanov N. A., Carlson H. A.: Exploring the composition of protein-ligand binding sites on a large scale. *PLoS Computational Biology 9*, 11 (2013), e1003321. doi:10.1371/journal.pcbi.1003321. 2

[KKF*17] Kozlíková B., Krone M., Falk M., Lindow N., Baaden M., Baum D., Viola I., Parulek J., Hege H.-C.: Visualization of biomolecular structures: State of the art revisited. In *Computer Graphics Forum* (2017), vol. 36, Wiley Online Library, pp. 178–204. 2

[KKL*16] Krone M., Kozlikova B., Lindow N., Baaden M., Baum D., Parulek J., Hege H.-C., Viola I.: Visual analysis of biomolecular cavities: State of the art. In *Computer Graphics Forum* (2016), vol. 35, Wiley Online Library, pp. 527–551. 2

[KSB*09] Krzywinski M., Schein J., Birol I., Connors J., Gascoyne R., Horsman D., Jones S. J., Marra M. A.: Circos: an information aesthetic for comparative genomics. *Genome research 19*, 9 (2009), 1639–1645. 3

[KSS*14] Kozlikova B., Sebestova E., Sustr V., Brezovsky J., Strnad O., Daniel L., Bednar D., Pavelka A., Manak M., Bezdeka M., et al.: Caver analyst 1.0: graphic tool for interactive visualization and analysis of tunnels and channels in protein structures. *Bioinformatics 30*, 18 (2014), 2684–2685. 2

[LMAPV15] Le Muzic M., Autin L., Parulek J., Viola I.: cellview: a tool for illustrative and multi-scale rendering of large biomolecular datasets. In *VCBM* (2015), pp. 61–70. 2

[LS11] Laskowski R. A., Swindells M. B.: Ligplot+: multiple ligand–protein interaction diagrams for drug discovery. *Journal of chemical information and modeling 51*, 10 (2011), 2778–2786. 3

[LW99] Liu M., Wang S.: Mcdock: a monte carlo simulation approach to the molecular docking problem. *Journal of computer-aided molecular design 13*, 5 (1999), 435–451. 1

[MI08] Momma K., Izumi F.: VESTA: a three-dimensional visualization system for electronic and structural analysis. *Journal of Applied Crystallography 41*, 3 (2008), 653–658. 2

[MI11] Momma K., Izumi F.: VESTA 3 for three-dimensional visualization of crystal, volumetric and morphology data. *Journal of Applied Crystallography 44*, 6 (2011), 1272–1276. 2

[MJD*16] Manguy J., Jehl P., Dillon E. T., Davey N. E., Shields D. C., Holton T. A.: Peptigram: a web-based application for peptidomics data visualization. *Journal of Proteome Research* (2016). 3

[SAMG14] Sarikaya A., Albers D., Mitchell J., Gleicher M.: Visualizing validation of protein surface classifiers. *Computer Graphics Forum 33*, 3 (2014), 171–180. doi:10.1111/cgf.12373. 3

[Sch16] Schrödinger L.: Schrödinger release 2016-1: Maestro version 10.5. lhttp://gts.sourceforge.net/, 2016. 3

[SEJ*05] Singh R. K., Ethayathulla A., Jabeen T., Sharma S., Kaur P., Singh T. P.: Aspirin induces its anti-inflammatory effects through its specific binding to phospholipase a2: Crystal structure of the complex formed between phospholipase a2 and aspirin at 1.9 å resolution. *Journal of drug targeting 13*, 2 (2005), 113–119. 7, 8

[SHL*12] Schneider N., Hindle S., Lange G., Klein R., Albrecht J., Briem H., Beyer K., Claussen H., Gastreich M., Lemmen C., et al.: Substantial improvements in large-scale redocking and screening using the novel hyde scoring function. *Journal of computer-aided molecular design 26*, 6 (2012), 701–723. 2

[SKR*14] Scharnowski K., Krone M., Reina G., Kulschewski T., Pleiss J., Ertl T.: Comparative visualization of molecular surfaces using deformable models. *Computer Graphics Forum 33*, 3 (2014), 191–200. doi:10.1111/cgf.12375. 3

[SMR06] Stierand K., Maass P. C., Rarey M.: Molecular complexes at a glance: automated generation of two-dimensional complex diagrams. *Bioinformatics 22*, 14 (2006), 1710–1716. 3

[SSH*15] Salentin S., Schreiber S., Haupt V. J., Adasme M. F., Schroeder M.: Plip: fully automated protein–ligand interaction profiler. *Nucleic acids research 43*, W1 (2015), W443–W447. 3

[SVGR16] Skanberg R., Vazquez P., Guallar V., Ropinski T.: Real-time molecular visualization supporting diffuse interreflections and ambient occlusion. *IEEE transactions on visualization and computer graphics 22*, 1 (01 2016), 718–727. doi:10.1109/TVCG.2015.2467293. 2

[TWWS98] Tanenbaum D. M., Wang Y., Williams S. P., Sigler P. B.: Crystallographic comparison of the estrogen and progesterone receptor?s ligand binding domains. *Proceedings of the National Academy of Sciences 95*, 11 (1998), 5998–6003. 7

[WL05] Wolber G., Langer T.: Ligandscout: 3-d pharmacophores derived from protein-bound ligands and their use as virtual screening filters. *Journal of chemical information and modeling 45*, 1 (2005), 160–169. 3

[WLRR99] Word J. M., Lovell S. C., Richardson J. S., Richardson D. C.: Asparagine and glutamine: using hydrogen atom contacts in the choice of side-chain amide orientation. *Journal of molecular biology 285*, 4 (1999), 1735–1747. 3

[ZMD13] Zhang H., Meltzer P., Davis S.: Rcircos: an r package for circos 2d track plots. *BMC bioinformatics 14*, 1 (2013), 244. 3