# InVADo: Interactive Visual Analysis of Molecular Docking Data

Marco Schäfer , *Member, IEEE*, Nicolas Brich , *Graduate Student Member, IEEE*, Jan Byška ,
Sérgio M. Marques , David Bednář , Philipp Thiel , Barbora Kozlíková , *Member, IEEE*,
and Michael Krone , *Member, IEEE*

*Abstract*—Molecular docking is a key technique in various fields like structural biology, medicinal chemistry, and biotechnology. It is widely used for virtual screening during drug discovery, computer-assisted drug design, and protein engineering. A general molecular docking process consists of the target and ligand selection, their preparation, and the docking process itself, followed by the evaluation of the results. However, the most commonly used docking software provides no or very basic evaluation possibilities. Scripting and external molecular viewers are often used, which are not designed for an efficient analysis of docking results. Therefore, we developed InVADo, a comprehensive interactive visual analysis tool for large docking data. It consists of multiple linked 2D and 3D views. It filters and spatially clusters the data, and enriches it with post-docking analysis results of protein-ligand interactions and functional groups, to enable well-founded decision-making. In an exemplary case study, domain experts confirmed that InVADo facilitates and accelerates the analysis workflow. They rated it as a convenient, comprehensive, and feature-rich tool, especially useful for virtual screening.

*Index Terms*—Molecular docking, AutoDock, virtual screening, visual analysis, visualization, clustering, protein-ligand interaction.

## I. INTRODUCTION

**M**OLECULAR docking has become an important technique in structural biology and computer-aided drug discovery. The goal of molecular docking is to predict the orientation and binding affinity of a ligand in the binding site of the target protein. In biochemistry, a ligand is usually a small molecule that can bind to a so-called target or receptor molecule—often a protein—to perform a specific biological function (e.g., activating or inhibiting an enzymatic reaction). Docking has been used for more than four decades and still plays an essential role in high-throughput virtual screening, especially in drug discovery [1], [2]. Ligand-target binding characteristics are investigated to find geometrically and chemically optimal fits, i.e., with a high binding affinity. For this, individual poses are identified using stochastic search algorithms and are evaluated by scoring functions that enable ranking of possible leads [3]. A binding pose is defined by the combination of the conformation and the orientation of a ligand [4]. The general procedure consists of the target and ligand selection, their preparation, and the docking calculations, followed by the evaluation of the results [5]. There are many commonly used docking tools as *AutoDock Vina*, *Gold*, *DOCK*, *MOE*, *Glide*, *rDock* (see [6] and Section II-A). These tools offer the users only very basic visualizations when evaluating the results. Therefore, freely available molecular viewers, such as *PyMOL* [7], *Chimera* [8], or *VMD* [9], are often used for visual inspection. However, they are also not specifically designed to support an efficient in-depth analysis of docking results. The typical analysis workflow includes processing the data in different tools and scripts, which is tedious and time-consuming.

Therefore, we designed an application to facilitate a comprehensive analysis of docking results by supporting and leading the users through the evaluation process. It structures and visualizes the data in multiple ways, e.g., by clustering, building various interactive tables, and multiple linked 3D and 2D visualizations, as well as by enriching the data with additional post-docking analysis results. This enables the user to make well-founded decisions on the docking results, e.g., extract drug candidates (lead compounds) or identify hot-spots for protein engineering. Our application was designed in close collaboration with domain experts by first collecting common tasks and inferring requirements and then refining it in an iterative process.

Our primary contribution is *InVADo* (Interactive Visual Analysis of Molecular Docking Data), a visual analysis application specifically designed to support the exploratory analysis of molecular docking data (see Fig. 1). It filters the docking data by docking scores and spatially clusters the remaining ligands. These clusters are the entry point for users to explore, analyze, and evaluate the docking results. Clusters can be selected in the

Fig. 1. *InVADo Application Overview*: ① *Docking Overview*: Visualizes the results of the clustered docking as a stacked bar chart combined with a box plot, where the bars represent the clusters. ② *Statistics View*: Gives cluster-specific information about ligands and their statistics in a tabular view with filter and sort options. ③ *Functional Groups View*: Provides cluster-specific information about functional groups (chemical substructures) of the ligands presented as an expandable and selectable treeview. It also extends the filter options of the *Statistics View*. ④ *Ligand View*: This table shows the individual binding poses offering ligand-specific details like the docking score and interactions. It is combined with a treeview presenting the functional groups of the current ligand. ⑤ *3D Visualization*: Offers interactive visualization of the clustered docking results with a *Radial Menu* to browse them. ⑥ *Sidebar Menu*: Collapsible control panel allowing users to adjust the appearance of the *3D Visualization* and to parametrize the clustering of the ligand binding poses and the functional groups.

3D view or via multiple linked plots and tables, each with an increasing level of granularity—ranging from a whole cluster to individual ligands and their binding poses. InVADo also incorporates data from post-docking analyses that determine protein-ligand interactions and functional groups, which are used for filtering and sorting to enrich the analysis. This allows the users to get an overview as well as detailed information about individual ligands, which can lead to more profound insights into the properties of the receptor and its potential binding partners.

As an additional contribution, we present an exemplary case study that shows the capabilities of InVADo and evaluated it in collaboration with biochemical researchers on a virtual screening data set [10] using structured expert feedback sessions. The experts confirmed an improved and accelerated analysis of the docking results and a clear benefit from the extensive exploratory workflow, that led to the identification of general interaction features and hot-spots. Furthermore, they indicated that they were able to gain helpful insights from the enriched data.

## II. BIOCHEMICAL BACKGROUND

In this section, we briefly review the necessary biochemical background that is needed throughout the paper. We start with docking events that typically happen at the so-called *binding sites* of a receptor molecule, e.g., a protein. Binding sites consist of a certain combination of amino acids that enable interaction and are somewhat exposed to external molecules. If a *ligand* (a small molecule) binds to this site, a certain biological function can be triggered (e.g., inhibiting or activating an enzymatic reaction). Ligands target macromolecules, for instance, receptor proteins for signal transduction, enzymes that catalyze reactions,

or the spike proteins of viruses [5]. Binding sites are often located in surface clefts or pockets, whereby lower-situated sites are connected to the outer environment through molecular tunnels [11]. Identifying and visualizing such pockets from the molecular structure is a challenging task, which has been extensively investigated [12]. However, the structural identification of pockets is not necessary in our case, as they will be inferred from the docking results. In the following section, we briefly introduce docking and the relevant physicochemical properties influencing the interactions between ligands and proteins. For an in-depth introduction to molecular biology, we refer to Lodish et al. [13] or Alberts et al. [14].

### A. Virtual Screening Using Molecular Docking

The general idea of virtual screening is to have a target/receptor molecule and dozens to hundreds of thousands of different ligands for which a search for a spatially and chemically optimal fit is performed. A fit is a ligand pose with a good docking score (high affinity) indicating that the ligand potentially binds well to the target. During the search process, many different conformations of a ligand are perturbed in the protein binding site or, in the case of blind docking, on the protein surface. These conformations of the ligand are called binding *poses* or *modes*. The docking result is a list of *ligand binding poses* that have the lowest free binding energy/lowest score (i.e., the energetically most favorable ligand-target complexes). Molecular docking can use various search algorithms and scoring functions to predict and rank binding poses. Although docking still has limitations regarding its accuracy, it is commonly used for the screening of large databases of ligands to narrow down possible drug candidates [1]. Due to the speed of modern docking algorithms and

computing hardware, this is much faster and more cost-effective than using wet lab experiments for the screening process.

Over the last 30 years, *AutoDock* [15] has been one of the most widely used docking tools and is still continuously improved [16], [17], [18]. It uses a Lamarckian genetic algorithm together with Monte Carlo simulations [3] to solve the docking task. Since molecular docking is a complex and extensive topic that is beyond the scope of our paper, we refer to the original publications for more information.

### B. Ligand Physicochemical Properties

Ligands have various specific physicochemical properties that can be used to infer their behavior in chemical reactions, their solubility, stability, and other properties. This is crucial to support the users in judging the ligands regarding their drug-likeness, identifying possible chemical modifications to increase affinity towards a docking target, or knowing how environmental conditions could influence the docking.

InVADo uses information from *ZINC 15* [19], a widely used ligand database containing more than 120 million drug-like compounds, for which it provides additional physicochemical properties. We use a selection of three relevant physicochemical properties that express the drug-likeness, which are partially based on the widely-used *"Lipinski's rule of five"* [20] that provides a likeliness of a drug being effective when taken orally by a human. The first one is the *molecular weight*, a basic property of the ligand that allows inferring its size. The second one is the octanol-water partition coefficient *log P*, which is also part of this rule. It describes the distribution equilibrium of a molecule between the aqueous phase and n-octanol greasy phase, indicating its hydrophilicity ($>0$ hydrophobic; $<0$ hydrophilic) [21]. The third one is called *fraction of $sp^3$* ($Fsp^3$) and is a further drug-likeness score that is the coefficient of the number of $sp^3$ hybridized carbon atoms and the total number of carbon atoms. $Sp^3$ hybridized atoms have four single bonds to other atoms, which tend to be uniformly distributed around the carbon in a tetrahedral shape. Small molecules with more $sp^3$ hybridized carbons are thus less planar, have increased solubility, and can occupy more target space, which makes them better drug candidates [22].

### C. Functional Groups & Interactions

For docking data enrichment, i.e., for supporting decision-making for protein engineers and ligand designers, it is necessary to determine functional groups and interactions. Functional groups are sets of atoms that have specific physicochemical properties [23], for example, a polar carboxylic, ether, amine, or hydroxyl group. To identify these groups, the tool *Checkmol* [24] can be used, which distinguishes between 204 different types. From those functional groups, it is possible to derive reaction or interaction partners, interactions, or possible modifications of a protein, making them crucial for molecular docking.

We consider hydrogen bonds, halogen bonds, hydrophobic interactions, metal complexes, $\pi$-cation interactions, $\pi$-stacks, and salt bridges as types of interaction. In docking, these are among the most important properties for the chemical fit of a

ligand to a given protein region. Together with their geometric fit, they lead to a lower or higher docking score. InVADo determines the mentioned interactions using the tool *Protein-Ligand Interaction Profiler (PLIP)* [25], which calculates non-covalent interactions at the atom level.

## III. RELATED WORK

Kozlíková et al. [26] gave a comprehensive overview of existing methods for the visualization of molecules. Similarly, Osolodkin et al. [27] provided a review of visualization techniques for the exploration of chemical space. Therefore, we only mention a few examples that are most relevant to our work. The data produced by molecular docking tools can be, to some extent, visualized in general molecular visualization tools, such as *PyMOL* [7], *VMD* [9], or *Chimera* [8]. Although some other tools provide means to visualize ligands (e.g., CaverAnalyst [28] or SAMSON Connect [29]), they are not designed to visualize results of molecular docking and, thus, some critical features are missing.

### A. Visualization of Molecular Interaction

Several research groups have focused specifically on the visual analysis of protein-ligand interactions. Furmanová et al. [30] proposed a method for the exploration of the geometric properties of the ligand transportation through the protein. Duran et al. [31] developed a system to explore long ligand trajectories via linked 2D and 3D views, offering enhanced charts that can be utilized for navigation to interesting parts of a simulation based on predefined properties. The methods proposed by Skånberg et al. [32] and Byška et al. [33] allow domain experts to define and extract their own properties. Schatz et al. [34] aggregate ligand trajectories on the protein surface to study typical paths to the binding site. However, all these methods are tailored for the analysis of molecular dynamics simulations and not molecular docking and, thus, cannot be easily extended to our case.

When analyzing docking data, it is crucial to understand the interactions between ligand atoms and protein amino acids. Hermosilla et al. [35] represented the interaction energies between a ligand and the surrounding amino acids with 2D and 3D arrows. As spatial representations often suffer from occlusion, alternative approaches provide such information with abstract representations. For example, LigPlot+ [36] uses a structure diagram to represent a ligand in detail while protein amino acids are abstracted to spoked arcs. Furmanová et al. [37] used similar structure diagrams, abstracting amino acids into stacked rectangles providing information about their properties. In both cases, interactions are indicated by lines connecting the ligand atoms and the amino acids. However, these methods cannot be utilized in our case, as they are designed for exploring a single protein-ligand conformation.

To explore multiple protein-ligand conformations over time, Vázquez et al. [38] proposed a 2D visualization that abstracts and aggregates the amino acids into a circular layout around the ligand representation. MolADI [39] allows for analyzing the evolution of the protein-ligand interactions over time using 2D plots. These visualizations can be used to analyze multiple

interactions over time, but both approaches are limited to a single ligand. Jurčík et al. [40] used a matrix of small plots to explore large ensembles of ligand trajectories. However, this work is also not directly applicable in our case as the focus is on the comparison of multiple trajectories of the same ligand.

### B. Large Ligand Ensembles

The most typical examples for the visualization of multiple ligands are tools such as ProteinPlus [41], which depicts the known ligands interacting with a selected protein directly in 3D or 2D using structure diagrams. ProteinPlus cannot be used in our case as it does not support the exploration of a large number of ligands or custom lists of docked ligands.

However, several tools have been proposed recently specifically for the exploration of large ligand ensembles. Janssen et al. [42] introduced a method that uses a scatter plot based on t-SNE embedding to compare the biological and chemical properties of a large number of ligands. A similar approach was presented by Sabando et al. [43], focusing on the comparison of different ligand groupings based on various descriptors. However, the scatter plots presented in these papers are not well suited for the exploration of the structural similarity of the ligands. To this end, Sabando et al. are using a separate 3D view for the spatial comparison of selected ligands. Gutlein et al. [44] used actual 3D models of ligands instead of points within the embedding.

Enhanced tabular views are the most common approach for exploring various properties of individual ligands within an ensemble. For example, Sabando et al. [43] based their solution on Taggle [45], while Data Warrior [46] heavily relies on the use of structure diagrams embedded directly in the tabular view to show the structural properties of individual ligands together with various quantitative values.

While all these aforementioned tools are well-suited for the exploration of large ligand ensembles, they are tailored to compare the ligands to each other, but they do not consider the protein-ligand interactions that are crucial in the case of molecular docking data. To solve this drawback, additional research is required, which is the main focus of this paper.

## IV. TASKS AND REQUIREMENTS

Our goal was to create an interactive visual analysis application for molecular docking data. The main idea is to guide the users to interesting ligands or hot-spots on the receptor molecule surface and to highlight over-represented physicochemical properties within those areas. This enables the users to investigate and answer questions about the receptor molecule and particularly well-suited ligands. Consequently, information extracted and derived from docking data is, for example, used in computer-aided drug design (virtual screening) or by protein engineers working on biotechnological applications (raw material production, e.g., fermentation, etc.). It also helps to answer important scientific questions, for instance, biologists researching signal transduction in organisms in which possible ligands for receptors are investigated. There is no clear *"standard"* workflow, because it depends on the specific research question

and data. In general, there is no coherent analysis environment, it is usually a combination of different tools and scripting (e.g., Excel, PyMOL [7], PLIP [25]), which leads to a complex and time-consuming workflow. Based on several discussions and semi-structured interviews with different domain experts working in pharmacology and biochemistry, we identified common tasks and usage scenarios to make InVADo as broadly applicable as possible.

The first task when analyzing docking results is typically to get an impression of the distribution of different ligands on the molecular surface. This includes inspecting ligands based on their free binding energies, sizes, energy distributions, functional groups, and more. That is, users need to browse the results, inspect, and filter them before performing further analysis steps.

A second task is to identify points of interest like clusters of chemical properties or specific ligands. This can indicate possible binding sites, which can be explored for their suitability as drug targets.

In addition to the investigation of clusters formed by ligands, users also need to determine the most frequent and highest-scoring ligands and explore their binding poses. These so-called *hits* are often possible drug candidates [47].

Besides finding the top-scoring hits, a more general and detailed analysis of all ligands is an often required task. This allows users to identify overall good binders, i.e., ligands that have high docking scores in all found clusters even if they are not the highest-scoring ones. The users are typically interested in identifying whether specific physicochemical properties or functional groups (called pharmacophores) are over-represented in a certain region of the protein. From this information, the affinity for a certain ligand type can be derived. It also allows drawing conclusions about the specificity of a certain binding site, i.e., inferring preferred binders for this binding site.

Similarly, it is important to assess the poses of the ligands, i.e., their spatial embedding in the context of interactions and chemical properties. This supports lead optimization. A *lead* is a ligand that is a drug candidate that can be further improved with small structural modifications [48].

To gain a deeper understanding of the potential interplay of a receptor and ligands, it can be necessary to enrich the docking results using additional data or analysis. One such task is the localization of suitable hot-spots for protein engineering, for instance, suitable locations for a mutation that leads to increased catalytic rates of enzymes, which is important in biotechnology. Additional data can also help to execute some of the preceding tasks. This includes interactions like hydrogen bonds or hydrophobic contacts as well as the previously mentioned chemical properties and functional groups. As these additional data are not part of the docking results their inclusion allows a more detailed evaluation of the docking.

Based on the tasks described above, we derived six requirements for our visualization application following the strategy by Brehmer and Munzner [49]:

R1 Provide a structured visual and textual access to the docking results, giving an overview that enables browsing and filtering as an entry point for further analysis.

R2 Identification of potential binding sites and other interesting hot-spots with high binding affinity.

R3 Hit identification by facilitating the determination of the most frequently docked and top-scoring ligands and binding poses.

R4 Support the identification of overall good binders and ligand types (not only top-scoring ones) for a comprehensive overview and specificity analysis.

R5 Visualization of docked ligands, interactions, and chemical properties to enable a detailed visual analysis.

R6 Enrichment of docking data to support evaluation of global docking results and decision-making.

## V. APPLICATION DESIGN AND FUNCTIONALITY

We designed our application InVADo to satisfy the requirements listed in Section IV. Due to the amount and complexity of the data and, consequently, the large number of required views, it is intended to be used on dual or ultra-widescreen monitors. InVADo is separated into two views (Fig. 1). The first one is a dashboard, which features plots and tables that give an overview, but also allows users to explore the data, e.g., to browse, search, and filter it, and to get additional details. The second one is a 3D view of the structure of the receptor molecule and the docked ligands, of the spatial location of the clusters and interactions, among others. It offers interactions like selection, filtering, or changing the molecular representation. Both views are tightly linked so that all selections and other changes are applied across all views. Below, we describe the design and functionality of both views as well as the data processing steps.

### A. Clustering and Binding Site Estimation

The first step is to identify locations of high binding affinity (R2), i.e., clusters of docked ligand binding poses. The goal is to identify hot-spots with a high density of docked poses while discarding the remaining binding poses that are not within one of these areas. Additionally, users can set a docking score threshold to discard poorly scored binding poses (R3/R4). Since the number of clusters is not known a priori, we opted for Density-Based Spatial Clustering (DBSCAN) [50]. The input parameters of DBSCAN are the maximum search radius and the minimum number of samples in a cluster. We use the centroids of the poses for clustering. We refer to poses that were discarded either by the score threshold or by not being in one of the DBSCAN clusters as *noise*. The clusters indicate possible binding sites. Thus, we extract the part of the molecular surface that is close to a cluster and mark it as a potential binding site.

### B. Dashboard

The design of our dashboard follows Keim's visual analytics mantra: *analyze first – show the important – zoom, filter and analyze further – details on demand* [51]. It consists of four views that feature an increasing amount of details about the docking results, the extracted clusters, the individual binding poses of a specific ligand, and the chemically relevant functional
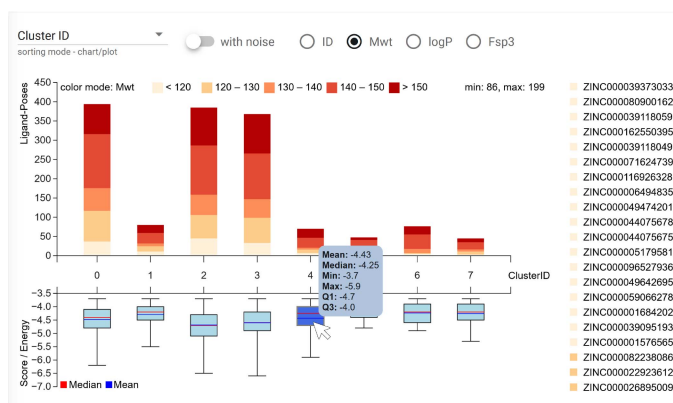


Fig. 2. *Docking Overview – Clusters*: a combination of a stacked bar chart and a box plot shows the ligand pose clusters. The bar chart visualizes the number of clusters and binding poses within a cluster. Quantized color maps are used to encode either the molecular weight (Mwt), the octanol-water partition coefficient (log$P$), the fraction of sp$^3$ (Fsp$^3$), or the ligand ID (see legend to the right). The box plot gives an overview of the docking score distribution within each cluster. Tooltips with more detailed information are available for both visualizations.

groups. This section describes these views and how they fulfill the requirements.

*1) Docking Overview – Stacked Bar Chart & Box Plot:* The docking scores and the extracted clusters are used to provide an overview of the docking results (R1). The clustered docking results are further aggregated and the information is presented as a stacked bar chart and a box plot (Figs. 1① & 2). Each cluster is represented by one vertically aligned bar and box plot, respectively. This combination allows the users to get an impression of how many ligands are within each cluster (height of the bars) and the distribution of docking scores within this cluster (box plot). The latter also contributes to fulfilling R3, as the plots can be used to identify extremes, e.g., the cluster with the highest amount of docked ligand poses, or the highest mean or absolute docking score. Note that the *AutoDock Vina* scores—which correspond to the docking energies—are negative, with smaller values signifying better docking.

Users can sort the clusters either by their ID, the number of ligand binding poses or according to the mean docking score of the clusters. Furthermore, they can choose to also show information about the poses previously classified as noise (see Section V-A). This will add a new bar and a box plot that contains the data of all noise poses.

Users can switch between four different coloring modes for the stacked bar chart (cf. Fig. 2): each ligand is assigned either an individual color (*ID*) showing how many binding poses of the same ligand are in one stack, or a coloring depending on one of three chemical properties is used (molecular weight *Mwt*, octanol-water partition coefficient *log P*, or fraction of sp$^3$ *Fsp$^3$*). This enables the users to see the chemical composition of the clusters—partially addressing R5—and allows them to discover similarities between ligands or clusters. As mentioned in Section II-B, the chemical properties are collected from the *ZINC* database or, if no ZINC name is present, are calculated with *PLIP* except *Fsp$^3$* (R6). To clearly show the value distribution of the chemical properties, the data is quantized into five bins,
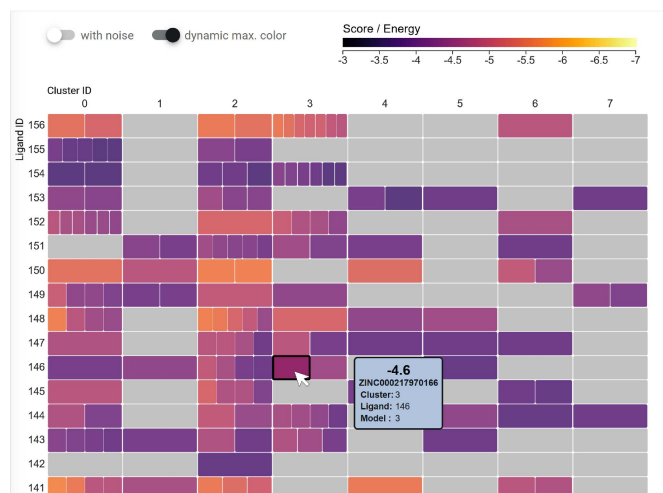
Fig. 3. *Docking Overview – Heatmap*: A segmented heatmap summarizes the docking and clustering results. Columns correspond to clusters and rows to individual ligands. Each cell of the heatmap where a cluster includes one or more poses of a ligand is drawn with one or multiple colored segments, depending on the number of binding poses present in this cluster. The segments are colored depending on their docking score using the inferno color map shown at the top. A tooltip with additional information is available for each segment (i.e., pose).



Fig. 4. *Statistics view:* Interactive table with cluster-specific information about the ligands. It provides the name, structural formula, min/max score values, and other statistics about each ligand in the currently selected cluster. The ligand in the table can be filtered by interaction types and searched for any given string or value.

to which a corresponding color is assigned. Note that only the value range between the quantile 5% and 95% is used, this is, we extended the first and last bin to reduce the influence of outliers and created equidistant bins in this range. When the cursor hovers over a stack, all stacks of the same ligand in other clusters are highlighted and a tooltip with detailed information about the pose, the ligand, and the values of the chemical properties appears. For the box plot, the tooltip shows the mean, median, box quantiles Q1/Q3, and the min/max values (see Fig. 2).

*2) Docking Overview – Heatmap:* As an alternative overview to the stacked bar chart and the box plot, we included a segmented heatmap (Fig. 3) that can be reached by switching to the HEATMAP tab (see Fig. 1①). The heatmap not only gives an overview of the docking results and the clustering (R1) but also specifically addresses requirement R4, as it allows the users to easily identify overall good binders. The columns of the heatmap are the clusters or binding sites, the rows are the ligands. We use a segmented heatmap, that is, if a ligand has multiple poses that belong to the same cluster, the corresponding rectangle of the heatmap is divided horizontally into multiple segments. That is, each segment represents a pose of a ligand. The segments are colored according to the docking score of the respective pose using the established *inferno* color map ( ), with black, mapped to the maximum value (poor binding; binding free energy/score closer to zero) and light yellow mapped to the minimum value (best binding; free energy/score strongly negative). Since the number of ligands usually exceeds the vertical screen space, the plot is scrollable. A tooltip shows the docking score, the *ZINC* name of the ligand, and the IDs of cluster, ligand, and pose. Similar to the bar charts and the box plot, the users can choose if only the clusters are shown or if the noise is displayed in an additional column of the heatmap.
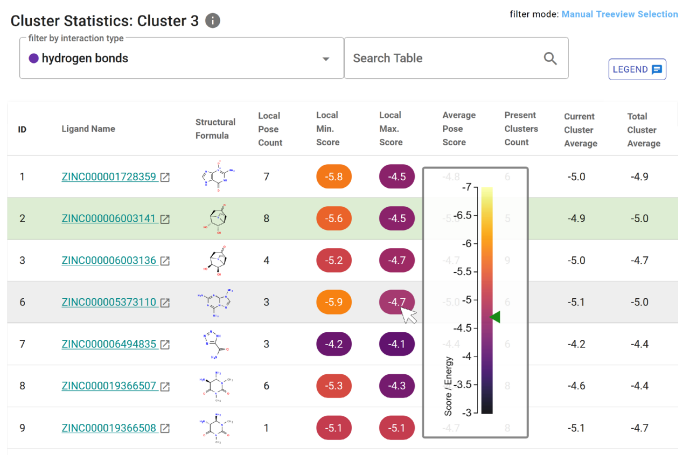
Each row can be seen as a docking profile that helps to identify overall good binders (R4). The heatmap directly shows whether the binding poses of a specific ligand are docked in many different binding sites or mainly one, and which poses reached high scores—shown by a "hot" color on the *inferno* color map—in any of the binding sites (R3). It also allows users to assess the identified binding sites (R2), e.g., the binding site specificity by checking if a binding site has only a few but good binders or many low-scoring ones.

*3) Statistics View:* The *Statistics View* (Figs. 1② & 4) is a table that provides detailed textual information and summary statistics about the ligands of a selected cluster or the whole docking data if no cluster was selected (R1). Clusters can be selected by clicking on the corresponding stacked bar or box plot. The table lists the name (i.e., the *ZINC* ID) and an image of the structural formula of all ligands included in the selected cluster, the number of docked poses of this ligand in the current cluster (*Local Pose Count*), the maximum and minimum scores of all poses in the cluster (*Local Min./Min. Score*), the average score of all poses of this ligand including non-clustered poses (*Avgerage Pose Score*) as well, how many clusters contain binding poses of this ligand (*Present Clusters Count*), as well as the average score for the poses of the selected cluster (*Current Cluster Average*) and the average score of all binding poses present in any cluster (*Total Cluster Average*). A tooltip provides more information about the currently hovered value. This can be either a color-coded scale of all docking scores with a marking for the score of the currently hovered pose (see Fig. 4), or a magnification of the structural formula for better readability. The table can be sorted by each column, which supports the user in identifying top-scoring ligands (R3). A ligand can be selected by clicking on the corresponding row highlighted in light green. Clicking on the name of a ligand in the second column opens the *ZINC* database webpage for this specific ligand, which provides more detailed information.
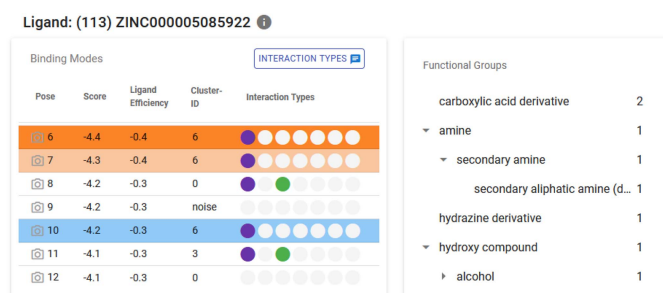
Fig. 5. *Ligand view*: This view provides information about the poses of a specific ligand using a table and a treeview. The table lists the docking score, ligand efficiency, cluster affiliation, and information about the presence of various protein-ligand interactions. The treeview to the right shows the hierarchy of functional groups and which ones are present in the currently selected ligand.
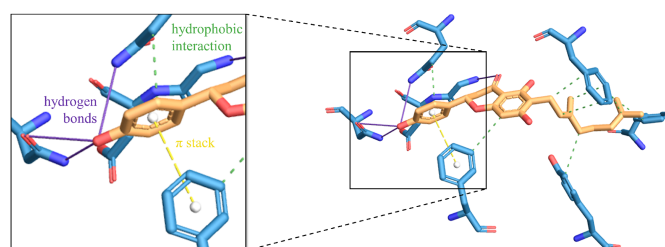


Fig. 6. *Protein-ligand interactions*: The flavonoid *Bonannione A* (orange) docked to *P-Gylcoprotein* (blue, only interacting protein residues are shown). Interactions are drawn as colored dashed or solid lines (yellow ●, green ●, purple ●). Image created by *PLIP* [25].

As the minimum and maximum scores for a ligand are important to identify good binders, we emphasize this information by coloring the background of these entries by the score value. We use the *inferno* color map (▬▬▬), with black mapped to the maximum value (poor binding) and light yellow, mapped to the minimum value (best binding).

The *Statistics View* also offers search and filter capabilities (R1). Users can filter the ligands using a drop-down menu to show only ligands that exhibit a specific interaction (see Section II-C). Furthermore, we included a text field to search for terms in all table columns, making it possible to search, e.g., for a specific *ZINC* name or a certain score.

*4) Functional Groups View:* As explained in Section II-C, functional groups are chemical building blocks of ligands that influence their binding behavior. That is, analyzing the presence or absence of certain functional groups is a crucial task in identifying good binders (R4). To facilitate this, we added a treeview that shows which functional groups are available either in the whole data set or in the currently selected cluster (Fig. 1③). Since docking data usually do not include explicit information about functional groups, we enrich the ligands with this information derived by *Checkmol* (R6). We chose a hierarchical view since similar functional groups can be assigned to super-groups. We used all groups listed for *Checkmol* to build their intrinsic hierarchy. The nodes of the treeview are expandable and the number of contained groups is shown to the right. For example, the hierarchy for *1,2-aminoalcohol* would be *hydroxy compound → alcohol → secondary alcohol → 1,2-aminoalcohol*. The treeview enables users to browse the functional groups, compare their number, and identify over-represented groups. This is interesting information, especially for ligand designers or protein engineers, e.g., regarding free binding energy optimization.

Groups can be selected using checkboxes, which act as a filter for the *Statistics View* table: ligands containing one of the selected functional groups are automatically detected, and their *ZINC* names are posted into the search field of the table.

*5) Ligand View:* After a cluster has been chosen from the *Docking Overview* and a ligand of interest is selected in the *Statistics View*, InVADo provides information about the ligand and its individual poses in the *Ligand View* (Figs. 1 ④ & 5). The data is again presented in a table that enables the user to

browse the pose properties and to compare and identify ligands with the highest ligand efficiency or certain interactions. Ligand efficiency is the binding free energy divided by the number of non-hydrogen atoms. The pose table has the same sorting options as the *Statistics View* to help identify good binders (R3, R4). It lists the score, ligand efficiency, and cluster assignment of each pose of the selected ligand. Furthermore, it shows which interaction types are present between the pose and the protein that are determined by *PLIP* and are presented as colored circles (R6). Based on discussions with domain experts, we integrated hydrogen bonds ●, halogen bonds ●, hydrophobic interactions ●, metal complexes ●, π-cation interactions ●, π-stacks ●, and salt bridges ●. The colors representing the interaction types follow the color conventions established in the domain and used by tools such as PLIP.

As shown in Fig. 5, some binding poses are automatically highlighted by a blue or orange background. Blue highlighted poses are in the selected cluster and orange ones additionally contain at least one of the currently selected functional groups (see Section V-B4). Clicking on a row will open a pre-rendered image of the protein generated by *PLIP*, showing a visualization of the selected pose of the ligand and the mentioned interactions (see Fig. 6).

We also show the functional groups of the ligand in a treeview to the right of the pose table. It is similar to the larger treeview in the *Functional Groups View* but only shows the ligand-specific functional groups.

### C. 3D Visualization

A fully interactive 3D visualization of the clustered docking results complements the dashboard (Fig. 1⑤). It can show various representations of the receptor protein, the ligands, interactions, functional groups, and more (Fig. 7).

Initially, only the protein and abstract representations of the clusters are visualized to give an overview. When selecting a cluster for detailed analysis, a *Radial Menu* appears that allows browsing the ligand binding poses within the cluster (Fig. 1⑤). Clustering, sorting, and filtering options can be adjusted and refined, and additional information can be added to the visualization to enhance the analysis. Users can apply a clip plane to reduce clutter and to better look into binding sites. The 3D view and the 2D views provided by the dashboard are tightly linked, that is, all selections and filters will be propagated to all
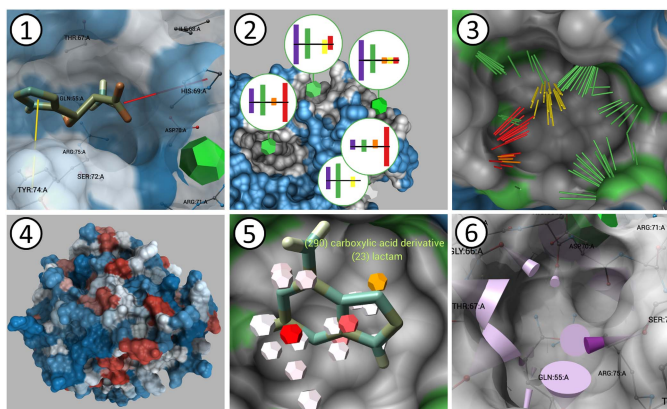
Fig. 7.  *3D visualization overview*: ① A *Ligand Pose* visualized as stick representation in a binding site. Surrounding protein residues are labeled and visible through the semitransparent SES. Interactions are shown as red and yellow sticks. ② *Interaction Type Bar Charts* summarize the interactions in a cluster. The top bars show the relative amount of interactions with respect to the binding site surface area and, the bottom bars with respect to the total interaction count in the cluster. ③ The *Interactions* in a binding site across all ligand poses visualized as sticks colored according to interaction type. Surface patches affected by hydrophobic interactions (green) are highlighted. ④ *Physicochemical Properties* (here hydrophobicity) can be color-mapped to the protein surface. ⑤ Clusters of *Functional Groups* are visualized as dodecahedra colored by the number of groups (white to red). On hover, a text label showing the contained groups is displayed. ⑥ *H-Bond Cones* summarize multiple hydrogen bonds between a protein atom and the docked ligands. The cone shows the average direction of the H-bonds and the opening angle distribution. The color indicates whether the protein is the acceptor (light purple) or the donor (dark purple).



Fig. 8.  Zoomed view of the *Radial Menu* showing the sub-circles. The center one shows the rank (or ID) of the ligand pose. To the right, the docking score and the presence of interactions are shown (here H-bonds). The outlined sub-circles to the left are control elements for which binding poses of a ligand are rendered: only the selected one, all poses of this ligand in this binding site, or all poses of this ligand.

other views to allow for a seamless analysis. The options of the 3D view can be adjusted via the sidebar menu of the dashboard (Fig. 1⑥).

*1) Protein and Cluster Visualization:* By default, the receptor protein is rendered using the Solvent Excluded Surface (SES), a smooth molecular surface that shows the interface between the protein and a specific small molecule like a solvent or ligand [26], [52]. It visualizes the 3D structure of the protein, which is an important aspect of docking. The spatial perception is enhanced using ambient occlusion [53], which particularly helps to get a better impression of the depth of cavities (i.e., possible binding sites; R2). The SES can be colored by the physicochemical properties of the protein, which helps to analyze the binding capabilities (R5). For example, the hydrophobicity coloring uses a diverging red-white-blue color map as shown in Fig. 7④. This coloring can support users to determine transmembrane proteins or hot-spots in a protein binding site that repel water and, thus, might be a good candidate for the interaction with certain ligands. The coloring mode and the color maps are user-adjustable parameters, including, e.g., coloring by element, B-factor, or protein chain.

The ligand clusters are visualized as dodecahedra placed at the centroid of the cluster (Fig. 7 ②) and colored using a linear color map from white to green (▭). The color expresses the quantitative difference in cluster size. We opted for dodecahedra instead of simpler spheres to visually differentiate them from the atoms, which are often represented as spheres. Selecting a cluster by clic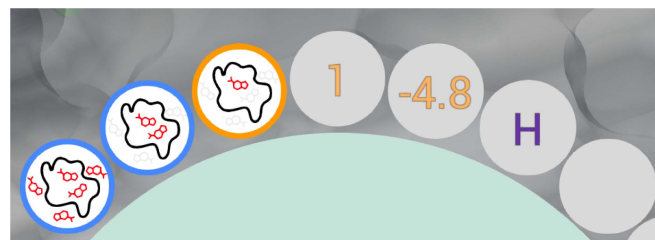king on the corresponding dodecahedron makes the dodecahedron disappear and the *Radial Menu* described below will appear.

*2) Radial Menu:* The *Radial Menu* (Fig. 1⑥) provides more information about the respective cluster. The circular layout encloses the point of interest and enables an efficient usage of the screen space. It allows the users to browse all docked ligands in the selected cluster (R1). Each of the circular menu items shows the structural formula of an individual ligand pose. The *Radial Menu* items are sorted by the score in descending order. Optionally, sorting can be extended by a second condition: the ligand pose must exhibit a certain interaction, otherwise it will get a rear position even if it has a high score. The number of menu items is user-adjustable. As typically not all results can be shown at once, the *Radial Menu* is divided into pages that can be browsed using the arrows above it. Above the arrows, a text label shows which ranks of ligand poses are currently presented (see Fig. 1⑤: ranks 1–8 out of 386 poses are shown, i.e., the user is on page 1 of 49).

Besides the rank and score of the current ligand pose, additional information is shown in sub-circles (Fig. 8). The small circles around each item show the rank (or ID) and the docking score of the pose, allowing users to compare binding poses and identify, e.g., the best-scoring ligands and poses (R3). To the right of the score are multiple sub-circles with fixed positions for each of the seven interaction types. If an interaction is present for the pose, colored characters indicating the interaction are drawn into the corresponding sub-circle (R6). If a menu item is selected, three further sub-circles appear to the left of the rank (see Fig. 8). Clicking on them allows the user to control which 3D models of that ligand are rendered (R1): by default, only the currently selected ligand pose is rendered (see Fig. 1⑤). This is indicated by an icon stylizing the protein binding site as a black line with only one small red ligand inside with the other ligands shown in light gray. The other options are to show all poses of the same ligand that are in the same cluster (indicated by the icon showing all ligands within the black outline in red) and to render all poses of that ligand regardless of cluster affiliation (indicated by red ligands inside and outside of the binding site).

*3) Ligand & Interaction Visualizations:* When a ligand is selected, e.g., via the *Radial Menu* or *Docking Overview*, it will be visualized as stick model colored by chemical element (Fig. 7①). In addition, the interactions between the ligand and

the protein can be visualized as thin, colored sticks, corresponding to and complementing the *Ligand View* of the dashboard (Section V-B5). This 3D visualization allows for a detailed spatial assessment of the docking pose and interactions with respect to the protein (R5). If a pose is selected, only the interactions present for this pose are shown. If no pose is selected, the user can opt to see all interactions of all binding poses. Depicting all interactions gives an overview of potential interactions, and their spatial and directional distribution (Fig. 7③).

Hydrogen bonds are usually the most prevalent interactions. Visualizing them as sticks can, thus, introduce visual clutter. To reduce this clutter while retaining information about the lengths, angles, and the spatial distribution of H-bonds that interact with the same protein atom, we aggregate them in *H-bond cones* (Fig. 7⑥). Each cone is oriented towards the average direction of all H-bonds and its length shows the average length of the H-bonds. The opening angle of the cone encodes the mean spatial distribution of the H-bonds. The cones are colored according to their type, i.e., whether the protein atom is the acceptor ▪ or donor ▪.

Surface patches corresponding to protein atoms that are affected by a certain interaction can be highlighted in the interaction-specific color (Fig. 7③). This allows users to locate preferred regions for certain interactions (R2). Furthermore, the surface color can also reflect the number of interactions counted for the individual protein atoms using a linear color scale (▬▬▬). This points to hot-spots with a high binding affinity (R2), which is valuable information for protein engineers and ligand designers.

*4) Binding Site and Functional Group Visualization:* Besides the aforementioned surface coloring by interaction type or the number of interactions, we added further options to allow users to explore spatial aspects of the docking data. To visualize the estimated binding sites based on the clusters, we offer a transparent rendering mode where the whole protein surface that is not close to the currently selected cluster is rendered transparently. This reduces clutter and helps to focus on the binding site (R1/R2). To achieve this, the binding site patches have to be derived from the ligands in the clusters. This is done by a neighbor search for each atom of all ligands using a distance criterion of 1.0 Å to find nearby protein atoms. Based on this, the parts of the surface corresponding to these found protein atoms can be determined (gray surface parts in Fig. 7②).

The ligands within a cluster are also used to find clusters of functional groups previously determined by *Checkmol*. The functional groups are represented as smaller dodecahedrons when compared to the dodecahedrons representing ligand clusters, which are also colored by size using a linear color scale (▬▬▬), as shown in Fig. 7⑤. The clustering has user-adjustable parameters for the minimum amount of cluster members and for the search radius. The clustering is performed individually for each super-group of functional groups. In the second step, the found functional group clusters are aggregated to avoid intersections of the dodecahedrons if they are too close to each other. They can be accessed by clicking, which will show a label with the amounts and functional group types. This will also select the corresponding functional groups in the *Functional Groups*

*View* (Fig. 1③). Additionally, the *Statistics View* is filtered accordingly. While hovering a dodecahedron it is highlighted in orange and the remaining dodecahedrons are hidden until a selection is made or the hovering ends.

*5) Interaction Type Bar Chart:* A *Interaction Type Bar Chart* can be displayed for each cluster/binding site and summarizes the interactions. The presented and aggregated information provides the user with an additional cluster-specific overview (R1), helps to compare clusters/binding sites, and allows for the identification of binding sites that have an interesting composition or profile of interactions (R2). The glyph-like visualization shown in Fig. 7② consists of two bar charts.

The top bars show the number of residues that belong to the binding site and are affected by an interaction. The bottom part shows the area of the binding site that is affected by each interaction. This information is derived from the surface areas of the interacting residues. The space-efficient bidirectional layout allows for a qualitative comparison of the absolute number of interactions versus the relative interaction surface area.

## VI. ARCHITECTURE & IMPLEMENTATION DETAILS

This section provides the implementation details of InVADo. As mentioned in the introduction, it is intended to be used in a dual or ultra-widescreen monitor setup to provide sufficient screen space for the large amount and complexity of the data and, consequently, the large number of required views. We implemented it as a two-window application consisting of a dashboard and a 3D visualization. Its client-server architecture is presented in Fig. 9.

For the *3D Visualization*, we used the open-source visualization framework *MegaMol* [54], [55], which is tailored to visualizing large scientific data sets and, in particular, molecular dynamics data. It is implemented in C++ and uses OpenGL for 3D rendering. Due to its modular architecture, it supports the prototyping of new visualizations with low overhead. With its various built-in biomolecular visualization features, it forms the basis of InVADo.

The dashboard was implemented as a web app using the *Vuetify* framework, which is a Vue UI library reducing the effort of web development by providing many pre-build components [56]. The interactive visualizations are generated using *Data-Driven Documents (D3)*, which binds data to DOM/SVG elements and manipulates them based on the data for visualization [57].

The source code of InVADo is publicly available:
```
https://github.com/MarcoSchaeferT/
InVADo_setup.
```

### A. Data, Preprocessing & External Tools

The input data are *Protein Data Bank, Partial Charge(Q), Atom Type(T) (PDBQT)* files (e.g., the ligands from the database *ZINC* [19]). PDBQT is also the output format of *AutoDockTools* and is used to store a set of ligand conformations consisting of atom positions and a docking score (see example in the supplemental material, available online). InVADo creates additional data by calling the external tools shown as turquoise
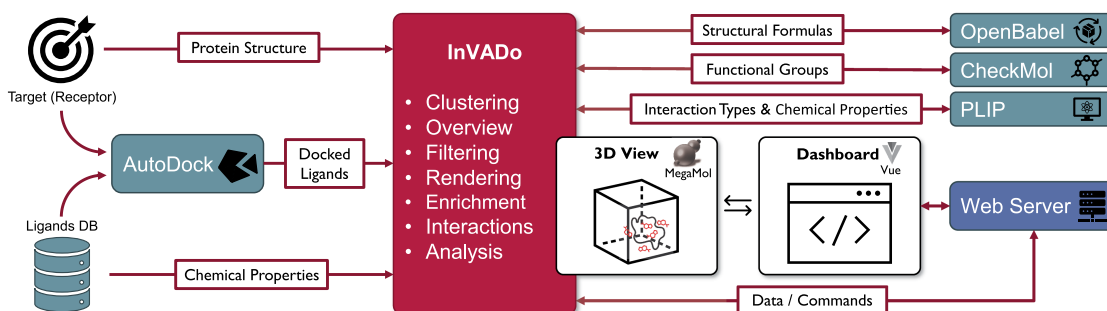
Fig. 9. *InVADo architecture:* The red rectangle at the center shows InVADo with some of its key features listed. Teal rectangles show external tools and data sources. To the left, the input data is listed, which consists primarily of the docking results, as computed by *AutoDock Vina*, the docked ligands from a database, and the target (or receptor protein) against which the ligands were docked. Additional data to enrich the docking data for a comprehensive visual analysis are provided by *PLIP*, *CheckMol*, and *OpenBabel*. InVADo controls all these external programs and automatically collects their output. All the data is processed, analyzed, and integrated by InVADo to produce an interactive 3D view (implemented using the MegaMol framework) of the clustered docking results. This view is complemented by a highly linked dashboard (implemented using the Vue library). Communication and data exchange between the two components is established via an additional Python web server.

boxes on the right side of Fig. 9. *OpenBabel* [58] creates the structural formulas of the ligands used in both the 3D and 2D views and converts the input data into the file format used by CheckMol. *CheckMol* [24] determines the functional groups of the ligands. *PLIP* [25] calculates the interactions for the protein-ligand complexes that are created by InVADo for each ligand pose.

To achieve high performance, InVADo is mainly written in C++. For smaller tasks related to data download, conversion, and communication with external tools, parallelized Python scripts controlled by the C++ program are used.

### B. GPU Accelerated Tasks & Rendering

InVADo requires neighbor searches for multiple tasks such as clustering, determining binding site patches or interacting residues of the protein surface, and the identification of residues adjacent to functional group clusters. Therefore, we use a CUDA implementation of the fast fixed-radius nearest neighbor search by Hoetzlein [59]. It significantly accelerated our implementation of DBSCAN [50], which we use to cluster the ligand poses and functional groups.

For fast rendering, we use GLSL shaders for raycasting spheres, cylinders, and cones [60]. We also use a modified version of the approximate ambient occlusion by Grottel et al. [53] to improve the spatial perception of the protein surface mesh. The CUDA implementation of the Thrust library is used to perform the view-dependent sorting of triangles necessary for transparent rendering of the surface.

## VII. RESULTS AND DISCUSSION

To evaluate InVADo, we performed a case study with a docking data set consisting of SARS-CoV-2 spike protein docked with FDA drugs [61]. Additionally, we collected expert feedback from biochemists in structured feedback sessions. In these sessions, the domain experts worked with an academically published data set collected for the analysis of P-Glycoprotein inhibitors [10].

### A. Case Study: SARS-CoV-2 Spike Protein Targeting

To create the data for this case study, we followed the approach of Pande et al. [61], who docked FDA drugs against the SARS-CoV-2 spike protein. The FDA drugs retrieved from *ZINC* did not include all ligands named in the paper but six of the nine top-scored were present (Ergotamine, Ponatinib, Yaz, Naldemedine, Conivaptan, Orap). We prepared the protein and used the drug data, in contrast to Pande et al. without further calculation steps like adding hydrogens or assigning new partial charges. The drug data were docked with *AutoDock Vina* (docking setup similar to Pande et al. [61]. Ten different ligand poses were calculated for each of the 2,215 ligands (paper: 1,565). InVADo was set up to cluster the docking results of 22,150 ligand binding poses with a minimum cluster size of 200, a search distance of 4.0 Å and a docking score of -6.0 kcal/mol.

The results as visualized by the combined stacked bar chart and box plot of the *Docking Overview* showed eight clusters with two similar-sized main clusters (see supplementary material for figures, available online). These two possible binding sites, in which 1,378 and 1,156 poses docked are located in the receptor binding domain (RBD) (amino acids: 319-541) reported by the paper, showing the validity of our clustering approach. Using the *ZINC*-IDs of top-scoring ligands from the publication as a search query for the *Statistics View* revealed that their poses are in an energy range from $-8.6$ to $-7.8$ kcal/mol, deviating from the reported range of $-8.2$ to $-6.5$ kcal/mol. This deviation can be explained by the missing ligand preparation, and that our data contains only a subset of six of the top-scoring nine ligands found in the publication. That is, InVADo helped us to verify that the differences are rather small, showing that our results are in line with the ones of the publication. With the help of the *Ligand View* table, we were able to determine that all six mentioned ligands have their top-scoring binding poses in the same binding site, i.e., the cluster coinciding with the RBD. This matches the results of Pande et al. in which these poses are located in the RBD as well.

InVADo offers many additional possibilities for detailed analyses of the docking results. The general summary of interactions presented by the *Interaction Type Bar Chart* showed

that H-bonds and hydrophobic interactions are the most prevalent interaction types (see supplementary material for figures, available online). This is in line with the findings of Pande et al. [61], who described that H-bonds *"play a crucial role in binding affinity, selectivity, and the stability"* of a protein-ligand complex and that the *"hydrophobic interactions also stabilize the complex"*. This can also be observed in the 3D view using the stick representation of the interactions or the interaction type-based surface coloring. In addition, the 3D view reveals detailed information about the locations of H-bonds and hydrophilic interactions in the binding site. The analysis of functional group clusters using the *Functional Groups View* showed that heterocyclic compounds and aromatic compounds are highly over-represented, both of which can be responsible for hydrophobic interactions. Amine- and hydroxyl-compounds, that can form H-bonds, are over-represented as well, although to a lesser degree. This further supports the findings of the *Interaction Type Bar Chart* and indicates a strong interaction in that binding site, which can be interesting for designing more specific drugs against SARS-CoV-2. To summarize, InVADo guides the users to the known RBD using the detected main clusters and moreover, it supports an in-depth analysis of the binding conditions, for instance, based on the aggregated interaction types and the functional group clusters.

### B. Expert Feedback: P-Glycoprotein Inhibitors

We evaluated InVADo with five domain experts (E1–5), who were asked to solve four different example tasks in the structured feedback sessions (see supplementary material for detailed task description, available online). The domain experts are biochemists and the mentioned tasks were created in consultation with one of them and one additional biochemist. We specifically included one of the experts (E5) who had helped us define the tasks and requirements to make sure that these were properly implemented in the final application. Due to their help in defining the user tasks, providing the screening data from their paper for the sessions, and their feedback on the initial design of InVADo, they are a coauthor of this paper.

The data used for our study was obtained as part of a previous project on the *Screening of Natural Compounds as P-Glycoprotein inhibitors against Multidrug Resistance* in the context of cancer treatment [10]. P-Glycoprotein is a pump that ejects foreign substances from cells (see Fig. 10). It is co-responsible for multidrug resistance in the context of using chemotherapeutics to fight cancer. The goal was to find inhibitors of the P-Glycoprotein as a treatment for multidrug resistance, so that the chemotherapeutics can work as intended before their efflux.

We asked the five biochemists to test InVADo and try to solve the four tasks. As mentioned above, four of them had not seen InVADo before. We provided a short video briefly explaining all the features of InVADo (see supplementary material, available online) and prepared a questionnaire. Furthermore, we captured the screen and audio of the test sessions. One of the experts was familiar with the data set and all of them work with molecular docking. Their self-assigned experience level ranges from beginner to proficient. The chosen data showcases the strengths of
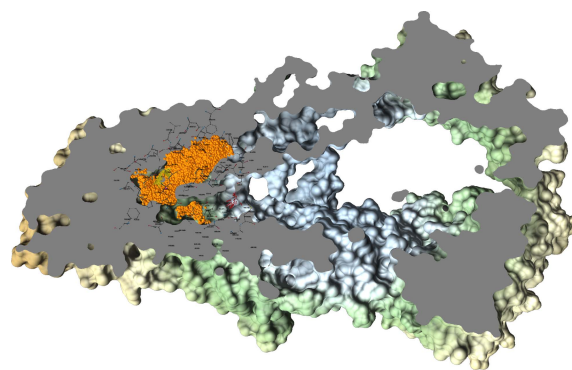


Fig. 10. Human P-Glycoprotein, a molecular pump for the transport of foreign substances out of the cell [10] (PDB ID: 4M1M), rendered in InVADo with a clip plane applied (gray). The orange spheres are ligand atoms of one of the clusters found by InVADo. The atoms are surrounded by the ball and stick representation of the interacting protein residues.

InVADo, as it contains a large protein for which all the preferred binding areas are located in the inner part of the protein. Thus, this data is more difficult to handle, and the domain experts have to make use of the clip plane to be able to see any binding pose or interaction (see Fig. 10). The data preparation and analysis workflow described in the original article [10] used a variant of *AutoDock Vina*, followed by the analysis of results by *MS Excel* and custom scripts. To visualize the results, the authors used *PyMOL*, for which they also needed to write further scripts. In contrast to this involved and time-consuming pipeline, our experts noted that InVADo is a well-designed, comprehensive tool with a nice, user-friendly interface (E2, E3, E4, E5). This confirms our claim that InVADo is an intuitive, easily accessible, and comprehensive tool for post-docking analysis.

Overall, all testers were able to solve the four tasks and the feedback was very positive. The collected feedback also contained helpful critiques and avenues for future development. We asked them to rate the following points on a five-point Likert scale ranging from *strongly dislike/disagree* to *strongly like/agree:* appearance, user interface design, application structure, offered visualizations, and interactivity. All categories reached an average of *strongly like* except *application structure*, which received an overall rating of *like* (see supplementary material for questionnaire results, available online).

We also asked if the experts would see a benefit of using InVADo compared to their current workflow and tools. On average, all of them *strongly agreed* that they would use and recommend InVADo for their future work. Only E1 noted that InVADo's application profile does not fit in with their currently used docking pipeline but added that they reckon it would be a *"wonderful tool for [...] virtual screening"*.

A general agreement among the experts was that the tool is overwhelming at first, as it offers *"a huge amount of options"* (E1). Specifically, they explained that the tool size makes it a bit hard to navigate if the options were not studied before, but *"there are other tools that are harder to learn"*. Furthermore, the expert stated that they realized that it is a general limitation of analysis programs to get all the information you want while simultaneously keeping it simple. They added that they needed a bit of explanation from us to solve all tasks in the intended

way. The testing time was not limited and on average it took ~70 minutes to explain the tool in more detail, explore any feature of InVADo, and solve the tasks. Nevertheless, the tool was rated as intuitive and "very user-friendly" (E4) with an average of *agree*. The domain experts said that the tool is very well-designed from the ligand perspective (E1, E3, E5). As a future extension, they suggested also adding further features from the protein perspective, e.g., the possibility to mark unwanted residues and automatically discard ligands interacting with these residues (E2). In addition, they were also interested to see the individual contribution of the interaction types (E5). This fits with the observation that the biochemists not only used the various overview features but also often started to make a detailed analysis of a single ligand binding pose to comprehend the presented interactions from the shown molecular structure of the ligand and the protein. This shows that InVADo is also suitable for the analysis of single ligand docking (*R5*). The fact that the domain experts reported that they would use InVADo also for pre- and post-optimization analysis further underpins it.

The experts liked the deep integration between the 3D view and the various panels. They described it as a convenient and comprehensive 3D view of ligands, clusters, and their functional groups, providing all the valuable information. They agreed that it allows getting a quick overview of how the binding works and mentioned that InVADo can be used to summarize the characteristics of binding with a "*very nice interface*" (E2) to arrange the different compounds. In this context, they also rated the offered tabular views as extremely useful and highlighted that they like the high number of features. Regarding the question of how InVADo fits into the existing pipeline, they mentioned that it would replace some steps of their pipeline, such as *MS Excel* and *PyMOL* (E3). Moreover, they usually have to do a lot of manual scripting to extract the important information in *PyMOL* and the layout of InVADo is very simple compared to this process. They also mentioned that they liked the surface coloring by interaction count, the segmented heatmap, and the cone representation of H-bonds, which was a completely new representation to them. It can help to see whether a ligand binds more precisely to the binding site than others.

They especially liked that the tool is also intuitive from a chemistry perspective, thus fitting their mental model. The functional group clusters were specifically mentioned because this feature allows for "*pharmacophore mapping*" (E5), i.e., "*mapping the protein region based on the functional groups*", which they rated as very useful for drug design. Among other things, the possibility to render multiple binding poses of the same ligand was rated as very good, as it allows for directly seeing the preferential binding poses of a ligand.

## VIII. CONCLUSION & FUTURE WORK

We presented InVADo, a novel approach for the visual analysis of molecular docking data. It is designed to give intuitive, exploratory, and structured access to docking results. The data is enriched by post-docking analysis results—e.g., interactions and functional groups—to enable a comprehensive analysis. These combined data are visualized in an interactive dashboard and a 3D visualization, offering multiple highly linked views.

InVADo offers many options for filtering and spatial clustering. The clusters are entry points for the analysis. InVADo is structured into views of various detail levels starting from a cluster overview to more granular, detailed views. More precisely, starting with *Docking Overview* as a summary of the found clusters and continuing to *Statistics View*, which also can present cluster-specific information together with the *Functional Groups View*. The most granular hierarchy level is the *Ligand View* providing specific information about the individual binding poses. The *Segmented Heatmap* bridges all different hierarchy detail levels. Interactions and selections in one view affect all other views, including the 3D visualization.

We evaluated InVADo using structured feedback sessions with domain experts, who confirmed an accelerated analysis workflow and rated InVADo as a convenient, comprehensive tool, which is very useful for virtual screening, especially in the context of pre- and post-optimization analysis. Furthermore, they indicated that they want to use it in the future to replace currently tedious steps in their analysis pipeline. They also rated the enriched data as helpful, e.g., for drug development and that they were able to derive insights from it for protein/ligand engineering.

In the future, we plan to add the possibility of performing a re-scoring for a single ligand or a selection of ligand poses. This will be part of a more complex integration of support for guided lead optimization. This should be further improved by visual methods to better focus an area of interest, e.g., part of the protein surface, by reducing visual clutter. Besides this, we also plan to further improve the spatial perception of the 3D view to enable the domain experts to get a more refined idea of how a ligand is located in a binding site and how the interactions can be formed.

## REFERENCES

[1] G. Wang and W. Zhu, "Molecular docking for drug discovery and development: A widely used approach but far from perfect," *Future Med. Chem.*, vol. 8, no. 14, pp. 1707–1710, Sep. 2016.

[2] V. Salmaso and S. Moro, "Bridging molecular docking to molecular dynamics in exploring ligand-protein recognition process: An overview," *Front. Pharmacol.*, vol. 9, Aug. 2018, Art. no. 923.

[3] B. Banaganapalli et al., "Molecular docking," in *Essentials of Bioinformatics: Understanding Bioinformatics: Genes to Proteins*, vol. 1, N. A. Shaik, K. R. Hakeem, B. Banaganapalli, and R. Elango, Eds., Cham, Switzerland: Springer, 2019, pp. 335–353.

[4] P. H. M. Torres, A. C. R. Sodero, P. Jofily, and F. P. Silva-Jr, "Key topics in molecular docking for drug design," *Int. J. Mol. Sci.*, vol. 20, no. 18, Sep. 2019, Art. no. 4574.

[5] G. M. Morris and M. Lim-Wilby, "Molecular docking," in *Molecular Modeling of Proteins*, A. Kukol, Ed., Totowa, NJ, USA: Humana Press, 2008, pp. 365–382.

[6] B. Zhang, H. Li, K. Yu, and Z. Jin, "Molecular docking-based computational platform for high-throughput virtual screening," *CCF Trans. High Perform. Comput.*, vol. 4, pp. 63–74, Jan. 2022.

[7] S. Yuan, H. S. Chan, and Z. Hu, "Using PyMOL as a platform for computational drug design," *WIREs Comput. Mol. Sci.*, vol. 7, no. 2, 2017, Art. no. e1298.

[8] E. F. Pettersen et al., "UCSF Chimera–A visualization system for exploratory research and analysis," *J. Comput. Chem.*, vol. 25, no. 13, pp. 1605–1612, Oct. 2004.

[9] W. Humphrey, A. Dalke, and K. Schulten, "VMD: Visual molecular dynamics," *J. Mol. Graph.*, vol. 14, no. 1, pp. 33–38, 1996.

[10] S. M. Marques, L. Šupolíková, L. Molčanová, K. Šmejkal, D. Bednář, and I. Slaninová, "Screening of natural compounds as P-Glycoprotein inhibitors against multidrug resistance," *Biomedicines*, vol. 9, no. 4, Mar. 2021, Art. no. 357.

[11] R. G. Coleman and K. A. Sharp, "Protein pockets: Inventory, shape, and comparison," *J. Chem. Inf. Model.*, vol. 50, no. 4, pp. 589–603, Apr. 2010.

[12] M. Krone et al., "Visual analysis of biomolecular cavities: State of the art," *Comput. Graph. Forum*, vol. 35, no. 3, pp. 527–551, Jun. 2016.

[13] H. Lodish et al., *Molecular Cell Biology*, 6th ed. San Francisco, CA, USA: Freeman, 2007.

[14] B. Alberts et al., *Molecular Biology of the Cell*, 5th ed. New York, NY, USA: Taylor & Francis, 2007.

[15] G. M. Morris et al., "AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility," *J. Comput. Chem.*, vol. 30, no. 16, pp. 2785–2791, Dec. 2009.

[16] O. Trott and A. J. Olson, "AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading," *J. Comput. Chem.*, vol. 31, no. 2, pp. 455–461, 2010.

[17] D. S. Goodsell, M. F. Sanner, A. J. Olson, and S. Forli, "The AutoDock suite at 30," *Protein Sci.*, vol. 30, no. 1, pp. 31–43, 2021.

[18] J. Eberhardt, D. Santos-Martins, A. F. Tillack, and S. Forli, "AutoDock Vina 1.2.0: New docking methods, expanded force field, and python bindings," *J. Chem. Inf. Model.*, vol. 61, no. 8, pp. 3891–3898, Aug. 2021.

[19] T. Sterling and J. J. Irwin, "ZINC 15 – Ligand discovery for everyone," *J. Chem. Inf. Model.*, vol. 55, no. 11, pp. 2324–2337, Nov. 2015.

[20] C. A. Lipinski, F. Lombardo, B. W. Dominy, and P. J. Feeney, "Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings," *Adv. Drug Del. Rev.*, vol. 46, no. 1, pp. 3–26, 2001.

[21] J. M. Sangster, *Octanol-Water Partition Coefficients: Fundamentals and Physical Chemistry*. Hoboken, NJ, USA: Wiley, May 1997.

[22] W. Wei, S. Cherukupalli, L. Jing, X. Liu, and P. Zhan, "Fsp3: A new parameter for drug-likeness," *Drug Discov. Today*, vol. 25, no. 10, pp. 1839–1845, Oct. 2020.

[23] P. Muller, "Glossary of terms used in physical organic chemistry (IUPAC Recommendations 1994)," *Pure Appl. Chem.*, vol. 66, no. 5, pp. 1077–1184, Jan. 1994.

[24] N. Haider, "Functionality pattern matching as an efficient complementary structure/reaction search tool: An open-source approach," *Molecules*, vol. 15, no. 8, pp. 5079–5092, Aug. 2010.

[25] S. Salentin, S. Schreiber, V. J. Haupt, M. F. Adasme, and M. Schroeder, "PLIP: Fully automated protein–ligand interaction profiler," *Nucleic Acids Res.*, vol. 43, pp. W443–W447, Jul. 2015.

[26] B. Kozlíková et al., "Visualization of biomolecular structures: State of the art revisited," *Comput. Graph. Forum*, vol. 36, no. 8, pp. 178–204, 2017.

[27] D. I. Osolodkin, E. V. Radchenko, A. A. Orlov, A. E. Voronkov, V. A. Palyulin, and N. S. Zefirov, "Progress in visual representations of chemical space," *Expert Opin. Drug Discov.*, vol. 10, no. 9, pp. 959–973, 2015.

[28] A. Jurčík et al., "CAVER Analyst 2.0: Analysis and visualization of channels and tunnels in protein structures and molecular dynamics trajectories," *Bioinformatics*, vol. 34, no. 20, pp. 3586–3588, Oct. 2018.

[29] OneAngstrom, "SAMSON," Accessed: Dec. 5, 2023. [Online]. Available: https://www.oneangstrom.com/

[30] K. Furmanová et al., "Interactive exploration of ligand transportation through protein tunnels," *BMC Bioinf.*, vol. 18, no. 2, Feb. 2017, Art. no. 22.

[31] D. Duran, P. Hermosilla, T. Ropinski, B. Kozlíková, A. Vinacua, and P.-P. Vázquez, "Visualization of large molecular trajectories," *IEEE Trans. Vis. Comput. Graphics*, vol. 25, no. 1, pp. 987–996, Jan. 2019.

[32] R. Skånberg et al., "VIA-MD: Visual interactive analysis of molecular dynamics," in *Proc. Workshop Mol. Graph. Vis. Anal. Molecular Data*, The Eurographics Association, 2018.

[33] J. Byška, T. Trautner, S. Marques, J. Damborský, B. Kozlíková, and M. Waldner, "Analysis of long molecular dynamics simulations using interactive Focus+Context visualization," *Comput. Graph. Forum*, vol. 38, no. 3, pp. 441–453, 2019.

[34] K. Schatz et al., "Visual analysis of large-scale protein-ligand interaction data," *Comput. Graph. Forum*, vol. 40, no. 6, pp. 394–408, 2021.

[35] P. Hermosilla, J. Estrada, V. Guallar, T. Ropinski, A. Vinacua, and P.-P. Vazquez, "Physics-based visual characterization of molecular interaction forces," *IEEE Trans. Vis. Comput. Graphics*, vol. 23, no. 1, pp. 731–740, Jan. 2017.

[36] R. A. Laskowski and M. B. Swindells, "LigPlot+: Multiple ligand– protein interaction diagrams for drug discovery," *J. Chem. Inf. Model.*, vol. 51, no. 10, pp. 2778–2786, Oct. 2011.

[37] K. Furmanová et al., "DockVis: Visual analysis of molecular docking trajectories," *Comput. Graph. Forum*, vol. 39, no. 6, pp. 452–464, 2020.

[38] P. Vázquez, P. Hermosilla, V. Guallar, J. Estrada, and A. Vinacua, "Visual analysis of protein-ligand interactions," *Comput. Graph. Forum*, vol. 37, no. 3, pp. 391–402, 2018.

[39] B. Bai, R. Zou, H. C. S. Chan, H. Li, and S. Yuan, "MolADI: A web server for automatic analysis of protein–small molecule dynamic interactions," *Molecules*, vol. 26, no. 15, Jan. 2021, Art. no. 4625.

[40] A. Jurčík et al., "Visual analysis of ligand trajectories in molecular dynamics," in *Proc. IEEE Pacific Visualization Symp.*, 2019, pp. 212–221.

[41] R. Fährrolfes et al., "ProteinsPlus: A web portal for structure analysis of macromolecules," *Nucleic Acids Res.*, vol. 45, no. W1, pp. W337–W343, Jul. 2017.

[42] A. P. A. Janssen et al., "Drug discovery maps, a machine learning model that visualizes and predicts kinome–Inhibitor interaction landscapes," *J. Chem. Inf. Model.*, vol. 59, no. 3, pp. 1221–1229, Mar. 2019.

[43] M. V. Sabando et al., "ChemVA: Interactive visual analysis of chemical compound similarity in virtual screening," 2020, *arXiv: 2008.13150*.

[44] M. Gütlein, A. Karwath, and S. Kramer, "CheS-Mapper 2.0 for visual validation of (Q)SAR models," *J. Cheminformatics*, vol. 6, no. 1, Sep. 2014, Art. no. 41.

[45] K. Furmanov á et al., "Taggle: Combining overview and details in tabular data visualizations," *Inf. Vis.*, vol. 19, no. 2, pp. 114–136, Apr. 2020.

[46] T. Sander, J. Freyss, M. von Korff, and C. Rufener, "DataWarrior: An open-source program for chemistry aware data visualization and analysis," *J. Chem. Inf. Model.*, vol. 55, no. 2, pp. 460–473, Feb. 2015.

[47] L. G. Ferreira, R. N. Dos Santos, G. Oliva, and A. D. Andricopulo, "Molecular docking and structure-based drug design strategies," *Molecules*, vol. 20, no. 7, pp. 13 384–13 421, Jul. 2015.

[48] B. K. Shoichet, S. L. McGovern, B. Wei, and J. J. Irwin, "Lead discovery using molecular docking," *Curr. Opin. Chem. Biol.*, vol. 6, no. 4, pp. 439–446, Aug. 2002.

[49] M. Brehmer and T. Munzner, "A multi-level typology of abstract visualization tasks," *IEEE Trans. Vis. Comput. Graphics*, vol. 19, no. 12, pp. 2376–2385, Dec. 2013.

[50] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Proc. 2nd Int. Conf. Knowl. Discov. Data Mining*, 1996, pp. 226–231.

[51] D. A. Keim, F. Mansmann, J. Schneidewind, J. Thomas, and H. Ziegler, "Visual analytics: Scope and challenges," in *Visual Data Mining*, S. J. Simoff, M. H. Böhlen, and A. Mazeika, Eds., Berlin, Germany: Springer, 2008, pp. 76–90.

[52] M. Schäfer and M. Krone, "A massively parallel CUDA algorithm to compute and visualize the solvent excluded surface for dynamic molecular data," in *Proc. 2nd Workshop Mol. Graph. Vis. Anal. Mol. Data*, 2019, pp. 1–9.

[53] S. Grottel, M. Krone, K. Scharnowski, and T. Ertl, "Object-space ambient occlusion for molecular dynamics," in *Proc. IEEE Pacific Vis. Symp.*, 2012, pp. 209–216.

[54] S. Grottel, M. Krone, C. Müller, G. Reina, and T. Ertl, "MegaMol—A prototyping framework for particle-based visualization," *IEEE Trans. Vis. Comput. Graphics*, vol. 21, no. 2, pp. 201–214, Feb. 2015.

[55] P. Gralka et al., "MegaMol – A comprehensive prototyping framework for visualizations," *Eur. Phys. J. Special Top.*, vol. 227, no. 14, pp. 1817–1829, Mar. 2019.

[56] Vuetify — A material design framework for Vue.js. Accessed: Dec. 5, 2023. [Online]. Available: https://vuetifyjs.com/en/

[57] M. Bostock, V. Ogievetsky, and J. Heer, "D$^3$ data-driven documents," *IEEE Trans. Vis. Comput. Graphics*, vol. 17, no. 12, pp. 2301–2309, Dec. 2011.

[58] N. M. O'Boyle, M. Banck, C. A. James, C. Morley, T. Vandermeersch, and G. R. Hutchison, "Open babel: An open chemical toolbox," *J. Cheminformatics*, vol. 3, no. 1, Dec. 2011, Art. no. 33.

[59] R. C. Hoetzlein, "Fast fixed-radius nearest neighbors: Interactive million-particle fluids," in *Proc. GPU Technol. Conf.*, Santa Clara, CA, USA, 2014, Art. no. 2.

[60] G. Reina and T. Ertl, "Hardware-accelerated glyphs for mono- and dipoles in molecular dynamics visualization," in *Proc. 7th Joint Eurograph./IEEE VGTC Symp. Vis.*, 2005, pp. 177–182.

[61] M. Pande, D. Kundu, and R. Srivastava, "Drugs repurposing against SARS-CoV2 and the new variant B.1.1.7 (alpha strain) targeting the spike protein: Molecular docking and simulation studies," *Heliyon*, vol. 7, no. 8, Aug. 2021, Art. no. e07803.

**David Bednář** is the team leader of molecular modeling and bioinformatics with the Department of Experimental Biology and RECETOX, Masaryk University in the Czech Republic. The group combines the usage of theoretical methods with an understanding of protein function. Their aim is to use molecular modeling and bioinformatics to uncover the principles of enzymology and medicine, leading to the creation of cutting-edge bioinformatics tools for protein engineering and analysis.



**Marco Schäfer** (Member, IEEE) received the master's degree in biosystems technology/bioinformatics from the Technical University of Applied Sciences Wildau, Germany. He is currently working toward the PhD degree with the Big Data Visual Analytics in Life Sciences Group, University of Tübingen, Germany. His research interests include the analysis and visualization of biological and chemical data with a focus on protein-ligand interactions.

**Philipp Thiel** received the diploma in bioinformatics from the University of Tübingen and worked as a PhD student with the Chemical Genomics Centre of the Max-Planck-Society in Dortmund. He worked as research scientist and scientific software developer with a focus on computer-aided drug design before starting as coordinator of the Institute for Bioinformatics and Medical Informatics, University of Tübingen, Germany.



**Nicolas Brich** (Graduate Student Member, IEEE) received the master's degree in biology and the master's degree in bioinformatics from the University of Tübingen, Germany. He is currently working toward the PhD degree with the Big Data Visual Analytics in Life Sciences Group, University of Tübingen, Germany. His main research focus is the visualization of multivariate and time-dependent biomedical data.

**Barbora Kozlíková** (Member, IEEE) is an Associate Professor with the Faculty of Informatics, Masaryk University in Brno, Czech Republic. She is the head of the Visitlab Research Laboratory, specializing in the design of visualization and visual analysis methods and systems for diverse application fields, including biochemistry, medicine, and geography. She has published more than 70 research papers.



**Jan Byška** is an assistant professor with the Masaryk University in Brno, Czech Republic and a part-time associate professor with the University of Bergen, Norway. He is a member of the Visitlab Research Laboratory, where his work focuses mostly on various challenges in the field of visualization of molecular and time-dependent data.



**Sérgio M. Marques** is a senior post-doc researcher with the International Clinical Research Center (FNUSA-ICRC) and Masaryk University, Brno, Czech Republic. His main research interests include the computational modeling of enzymatic systems for potential biotechnological applications. With extensive experience in experimental chemistry and biochemistry, he has worked exclusively in the field of computational biology for the last 10 years. He has co-authored 56 articles, 1 book chapter, and 2 patents.



**Michael Krone** (Member, IEEE) received the PhD (Dr. rer. nat.) degree in computer science from the University of Stuttgart, Germany. He is a junior professor with the Department of Computer Science, University of Tübingen, Germany. He has been a visiting assistant professor with New York University and a guest lecturer at Ludwig Maximilian University of Munich. His research interests include visualization, computer graphics, and human-computer interaction, with a focus on visual analysis of molecular and biomedical data.