

Multiscale Visual Drilldown for the Analysis of Large Ensembles of Multi-Body Protein Complexes

Katarína Furmanová, Adam Jurčík, Barbora Kozlíková, Helwig Hauser, and Jan Byška

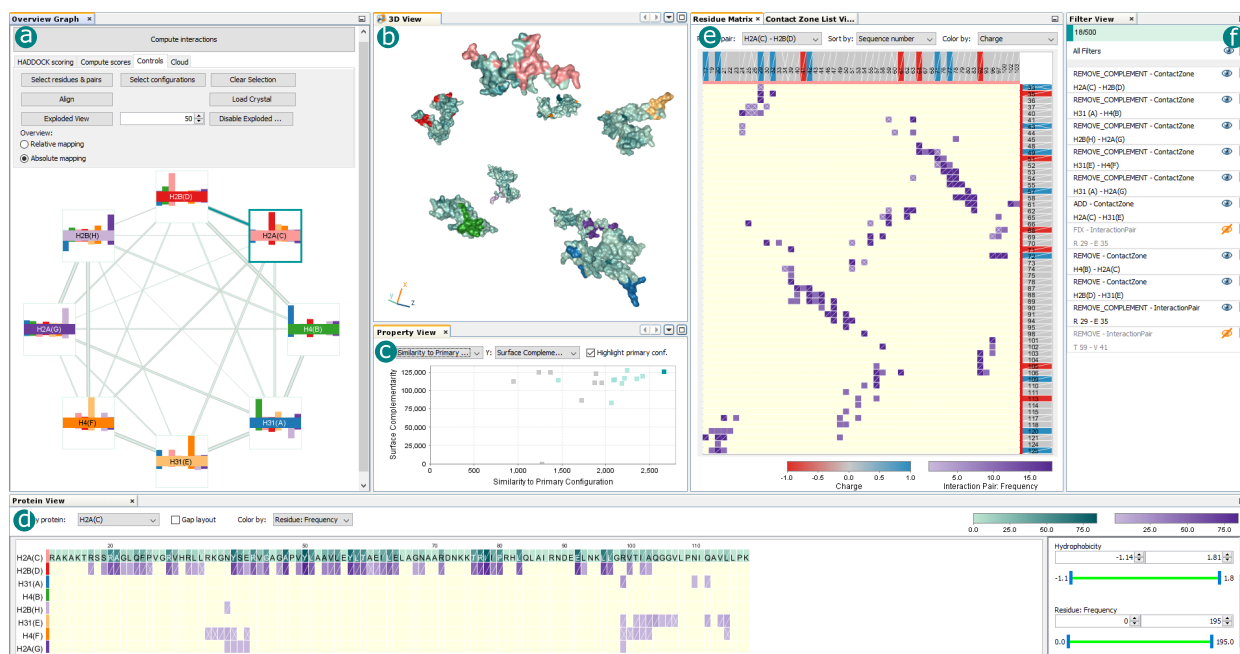


Fig. 1. Exploration of an ensemble of 500 possible human nucleosome configurations, based on the following components: a) Overview Graph with main application controls; b) 3D View, showing the exploded Complex Configuration, amino acids from contact interfaces are colored according to the frequency of their interactions; c) Property View, showing properties of all Complex Configurations; d) Protein View with range filter panel; e) Residue Matrix, which also can be switched to Contact Zone List View, and f) Filter View.

Abstract— When studying multi-body protein complexes, biochemists use computational tools that can suggest hundreds or thousands of their possible spatial configurations. However, it is not feasible to experimentally verify more than only a very small subset of them. In this paper, we propose a novel multiscale visual drilldown approach that was designed in tight collaboration with proteomic experts, enabling a systematic exploration of the configuration space. Our approach takes advantage of the hierarchical structure of the data – from the whole ensemble of protein complex configurations to the individual configurations, their contact interfaces, and the interacting amino acids. Our new solution is based on interactively linked 2D and 3D views for individual hierarchy levels. At each level, we offer a set of selection and filtering operations that enable the user to narrow down the number of configurations that need to be manually scrutinized. Furthermore, we offer a dedicated filter interface, which provides the users with an overview of the applied filtering operations and enables them to examine their impact on the explored ensemble. This way, we maintain the history of the exploration process and thus enable the user to return to an earlier point of the exploration. We demonstrate the effectiveness of our approach on two case studies conducted by collaborating proteomic experts.

Index Terms—Molecular visualization, data filtering, coordinated and multiple views.

1 INTRODUCTION

Protein complexes, formed by groups of associated polypeptide chains, facilitate a vast range of functions in living organisms, including the

- K. Furmanová, A. Jurčík, and B. Kozlíková are with Masaryk University; E-mail: katarina.furmanova@gmail.com, {xjurc, xkozlik}@ft.muni.cz
- H. Hauser is with the Univ. of Bergen; E-mail: Helwig.Hauser@UiB.no
- J. Byška is with the Univ. of Bergen and with Masaryk University; E-mail: jan.byska@gmail.com

Manuscript received xx xxx. 201x; accepted xx xxx. 201x. Date of Publication xx xxx. 201x; date of current version xx xxx. 201x. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org. Digital Object Identifier: xx.xxx/TVCG.201x.xxxxxx

organization of DNA and metabolic processes. Furthering the understanding of protein functions is thus essential for advances in many areas of medicine, biology, and chemistry [21, 28]. The function of a protein complex is tightly bound to its spatial structure and the interactions between individual protein units. Revealing the structure of a protein complex requires costly and time-consuming experiments. Computational tools, predicting the mutual interactions of proteins, are therefore increasingly employed to aid experimental processes.

While most of the computational tools are limited to the prediction of binary complexes, many of the protein complexes consist of multiple units [17]. To understand their function, it is necessary to take into account how all the units are interacting with each other. Irrespective of the computational approach, completely resolving the entire protein complex is a very costly combinatorial task, resulting in large ensembles of possible spatial arrangements of the protein complex.

Most of the multi-body docking tools, as well as several post-docking

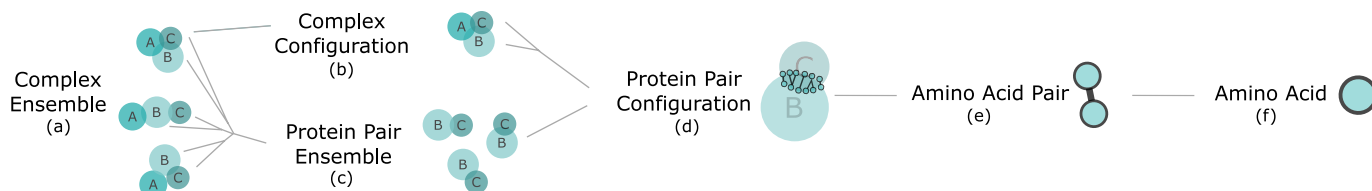


Fig. 2. Structure of a protein complex ensemble containing possible spatial configurations of a protein complex consisting of proteins A, B, and C.

tools, also evaluate the resulting configurations with a scoring function, taking into account, e.g., shape complementarity at the inter-molecular interface, energies (Van der Waals, electrostatics, etc.), or knowledge-based potentials derived from known evolutionary process (e.g., conservation) [6, 7]. The aim of the scoring function is to select the best representatives from the ensemble, containing thousands of docked configurations. Because of the large variety of the protein complexes and the differences between their structures and function, the tools and scoring functions are often tailored to a specific type of data and their usefulness for arbitrary protein complex is limited. This is also proven by ongoing CAPRI experiments [10] that focus on the assessment of performance of the docking and scoring tools for newly identified protein complexes.

As such, it is necessary for the domain experts to evaluate the results, assess the appropriateness of the scoring, and select the best representatives. Currently, this is done based on a cumbersome one-by-one exploration of the docked configurations in general molecular visualization tools such as PyMOL [27], or in tools dedicated to the exploration of pairs of interacting proteins such as CoCoMaps [31] or COZOID [4]. These tools are tailored to the evaluation of binary docks, if at all, and lack proper support for more intricate datasets.

Therefore, we propose a novel interactive visual drilldown approach to the exploration of multi-body protein docking results. We take advantage of the hierarchical nature of the data and propose a system enabling domain experts to explore, compare, and filter protein complexes at different levels of detail. We propose several dedicated views, displaying the available information for each level in the data hierarchy. As the views are interactively linked, the users can observe how a given filter or selection operation translates to other levels of the data hierarchy. We also track all filtering operations and provide an interface where the users can check how previously used filters affect the explored ensemble and, if necessary, revert them.

The bi-directional linking across several levels of such hierarchy is a challenging task as the propagation of the user’s actions through the levels of the hierarchy can lead to large changes in the individual views. As this can be confusing, we have to carefully consider how and when the user-triggered changes are propagated. As hierarchical data appear in many domains, the presented solution for multi-level filtering and selection is one of the main contributions of this paper.

2 TASKS AND DATA ANALYSIS

Our solution was designed in tight collaboration with proteomic experts from two research groups, one focusing on structural biology and one on protein engineering. We conducted several focus groups at the beginning of the design process as well as after the experts had a chance to test the first prototype. The main topic of the focus groups was the discussion of the current workflow and the identification of challenges connected to the analysis of multi-body protein complexes.

In the initial focus group we discovered that proteomic experts explore results from computational docking tools to identify the biochemically most relevant spatial arrangements of the examined protein complex that can be experimentally verified in the lab. When performing this analysis, the domain scientists draw heavily on existing knowledge from the literature. To be able to incorporate this knowledge during the analysis process, the proteomic experts need to be able to quickly explore and filter the predicted arrangements concerning a multitude of biochemical and spatial properties on different levels of detail. Without proper visual support, this is a very tedious and

time-consuming process as they need to explore large and intricate ensembles of predicted configurations.

We identified the following hierarchical structure of the data for protein complexes. We call the initial dataset produced by a computational docking tool a **Complex Ensemble** (CE, Figure 2a), containing the entire dataset of all computed configurations. It consists of many **Complex Configurations** (CC, Figure 2b), where each configuration represents one possible spatial arrangement of the whole protein complex. The Complex Configuration can be further split into individual **Protein Pair Configurations** (PPC, Figure 2d), each representing exactly one mutual position of two proteins from the protein complex. Each Protein Pair Configuration contains a set of **Amino Acid Pairs** (AAP, Figure 2e), each consisting of two interacting amino acids (AA, Figure 2f) from the two proteins forming the Protein Pair Configuration. Alternatively, we can split the Complex Ensemble into different **Protein Pair Ensembles** (PPE, Figure 2c), each consisting of the set of all PPCs that are formed by the same protein pair. In our solution, we employ run-time data structures that correspond one-to-one to the above-described structure. Such structure was sufficient to perform filtering and selections in real time across all scales in all tested datasets.

At each level of this hierarchy, we observe different properties, which can be explored and used for the identification of the biochemically most relevant CCs. Individual CCs and their PPCs can be scored based on energy measures between the interacting atoms (e.g., Van der Waals or electrostatic energies), where smaller numbers indicate more favorable interactions, or based on their geometrical properties (e.g., surface complementarity).

The docking tools often use a combination of these and other properties to define their own scores and select the most relevant representatives. However, these scoring functions can give different results for different protein complex types. Thus, the experts cannot fully rely on this ranking and filter out some solutions solely based on the scores of the computational tools.

In our solution, these properties and scores can be pre-loaded from the docking results, if provided (e.g., from HADDOCK [12]). Additionally, we include a property indicating the similarity of Complex Configurations (CC) and Protein Pair Configurations (PPC) to one selected, so-called primary CC or PPC. This measure is useful in cases when the domain expert already identified a potentially relevant Complex Configuration and searches for similar Complex Configurations, possibly revealing an even better solution. We also enable to compute this similarity with respect to a protein complex that is not part of the explored ensemble, to include the domain knowledge from partially resolved models or to compare results among different species. The similarity score computation is adapted from our previous work [4].

In terms of interacting amino acids, the domain experts are often interested in exploring the frequency of their occurrence in a dataset, distances between them, or their physico-chemical properties, such as hydrophobicity (i.e., tendency to interact with water molecules) or charge, both indicating the feasibility of a given interaction.

Based on the focus groups with domain experts, we identified a set of tasks that a visualization system for the analysis of large ensembles of multi-body protein complexes should support:

- T0** The primary goal is to identify the biochemically most relevant Complex Configurations from a large ensemble of predicted spatial arrangements.
- T1** To enable this, we need to provide the ability to explore all Complex Configurations and their spatial and biochemical properties, and support the identification of the interacting proteins.

- T2** It is also necessary to provide means for quick identification of potentially important Protein Pair Ensembles and the exploration of individual Protein Pair Configurations and their properties.
- T3** The solution has to enable the domain experts to identify the interactive Amino Acid Pairs and explore their properties both globally and separately for each Protein Pair Configuration.
- T4** In order to decrease the size of the exploration space, it is necessary to filter out invalid predictions based on their properties on all levels of the hierarchy as soon as they are discovered.
- T5** To exploit domain knowledge and ease the exploration, experts require the ability to compare the main differences between individual arrangements within the studied ensemble as well as differences to other protein complexes.
- T6** Furthermore, the system should provide information about the exploration steps that the domain experts undertook and enable the expert to adjust or revert them.

3 RELATED WORK

Our work is related to several research areas which we can generally divide into two main parts. The first part is dealing with computational approaches, producing ensembles of possible configurations. The second area covers the visual exploration of these ensembles, including the visual representation of protein complexes on different levels of detail and their visual analysis using drilldown and filtering.

3.1 Computation of Protein Complexes

Most of the currently available computational tools for protein-protein interactions are focusing on protein pairs. A comprehensible overview was published by Huang [6]. Some of the existing approaches, such as ArDock [25], already combine the computational method with a basic visual representation of the predictions. There are even some solutions, such as DockingShop [18], which are enabling the user to interactively design an initial configuration for a protein docking prediction process through a molecular graphics interface.

In the past decade, several tools emerged which enable the prediction of multi-body protein complexes. To compute a spatial arrangement of bigger assemblies, some computational tools reuse the pair-wise docking of the involved proteins and combine them once computed (e.g., [3, 32]), while others rely on restraints derived from experimental data to limit the exploration space (e.g., [12, 13]).

One of the first tools designed primarily for multi-body docking was CombDock [9]. The algorithm works on a principle of hierarchical construction of the complex from smaller subunits and a greedy selection of the best-ranking subunits. The combinatorial step is followed by the reduction of solutions based on RMSD and a scoring function. Multi-LZerD [3] uses a genetic algorithm to generate complexes from initial pairwise docks and applies an energy minimization structure refinement procedure for the ranking of the solutions. Venkatraman et al. [32] proposed an ant colony optimization approach to solve the combinatorial problem. DockStar [1] formulates the task of detecting the spatial conformation of a protein complex as an Integer Linear Program. Unlike other methods, it also integrates experimental data from mass spectrometry into the scoring of the solutions. Another tool reusing pairwise docks in combination with experimental data is PRISM-EM [13]. It uses density maps from cryo-electron microscopy for guiding the placement of subunits.

While all of the tools mentioned above build the complex structure from samples of pairwise prediction, HADDOCK [12], to the best of our knowledge, is the only tool that performs simultaneous docking of multi-body complexes. During the docking process, it uses constraints from different kinds of experimental data to drive the formation of the complex. However, unlike DockStar [1], which supports the prediction of complexes consisting of up to 16 units, HADDOCK allows predicting only complexes with up to 6 units.

3.2 Visualization of Protein Complexes

Protein complexes, irrespective of the number of units, can be in general visualized in one of the traditional and widely used software tools, such as UCSF Chimera [22], PyMol [27], VMD [8], Aquaria [20], CAVER

Analyst [11], and others. However, these tools provide the users only with general molecular representations, without specific support for the analysis of protein complexes and their visual exploration. This need was addressed by Lee and Varshney [16] by visualizing the volume of the area of the docking site between two proteins. Another attempt of a schematic visualization of two interacting proteins was proposed in the DIMPLOT extension of the LigPlot+ tool [15]. DIMPLOT shows the protein-protein interface in a 2D diagram. The 3D representation of this diagram can be viewed in PyMOL [27] or RasMol [23, 26]. A 2D schematic representation of interacting areas between proteins is available also in the PDBsum database [14].

Although these schematic representations are conveying the information about a single configuration, they do not support the comparison and interactive filtering of entire ensembles of configurations. This issue is addressed in the CoCoMaps [31] and COZOID [4] tools. Both tools come with linked visualizations, aiding the users in analyzing and comparing interactions between protein pairs. CoCoMaps and its successor CONS-COCOMAPS [30] enable to measure and visualize the consensus in multiple docking solutions and display the conservation of residue contacts using intermolecular contact maps. The COZOID tool uses a set of linked views for the interactive exploration of large ensembles of protein pairs, supporting a visual drilldown approach for narrowing down the set of possibly relevant configurations. The main limitation of these approaches is that they are operating only on protein pairs (i.e., single Protein Pair Ensemble) and cannot be directly applied to multi-body complexes. The multiscale aspect in molecular visualization can be explored on different granularity levels, as shown in the recent survey of Miao et al. [19].

In our case, we were not only concerned with designing proper visual representations of the individual hierarchy levels of large ensembles of multi-body complexes, but also with how to interactively explore and filter these ensembles to support the identification of biochemically most relevant instances. Splechtna et al. [29] focus on the problem of interactive visual steering of hierarchical simulation ensembles. In their substantially different application case, they also deal with linking representations on different levels of detail as well as with the challenge that the ensemble can grow during the exploration process.

4 METHOD OVERVIEW

The primary goal of this work was to enable biochemists to identify the biochemically most relevant spatial arrangements of multiple interacting proteins (**T0**) from a large ensemble of possible Complex Configurations (CC), predicted by a computational tool. To support this goal, we designed our solution as an interplay between several linked 2D and 3D views. Figure 1 shows an overview of the solution and all its components. The high-level overview of all loaded CCs is available in the Overview Graph (a) which serves for the preliminary filtering of CCs for further scrutinization. Details about selected properties of CCs of interest are visible in abstract views (c), (d), and (e), and the most detailed view of individual CCs is the 3D view. Here we provide an exploded view as the most interesting parts of the complex are the contact zones, hidden inside the complex. Due to the option of advanced filtering operations, we also added a specific panel supporting this task. In the following, we first describe the proposed interaction concept and then the individual views in more detail.

4.1 Interaction Design

The phenomena at hand have a natural hierarchical structure where all levels of the hierarchy have to be explored (**T1**, **T2**, **T3**) by domain scientists. The users can perform filtering and selection operations in each view to support the exploration of a large number of Complex Configurations (CC). The filtering operations define rules describing which CCs are irrelevant and thus should not be displayed to the user in subsequent analysis steps (unless the user specifically recalls them to change/reset the filtering) it is also interactively determined which CCs are important and should be kept. The selection operations, on the other hand, define the current focus of the user, allowing to highlight more important data while suppressing the context.

We elicited four primary filtering operations:

- **Remove:** filters all Complex Configurations satisfying a rule, i.e., all selected CCs or all CCs with a selected Amino Acid Pair.
- **Remove complement:** filters all Complex Configurations not satisfying a rule.
- **Fix:** prevents Complex Configurations from being removed from the ensemble by other filtering operations – this operation does not remove anything.
- **Add:** re-adds the previously filtered Complex Configurations back to the ensemble for further exploration.

When designing the filtering operations, we have considered merging the *remove complement* and *fix* operations into one. An exemplary case for such a merged operation would be the following: the user knows from experimental results that a certain pair of amino acids must interact in the protein complex. Thus, he or she would search for the given Amino Acid Pair and fix it and remove all Complex Configurations that do not contain this pair. However, this would effectively block the user from any further drilldown, since all the Complex Configurations containing this Amino Acid Pair (potentially still a large number) would be fixed. Therefore, we decided to separate the operations. While the *remove complement* operation is useful in the above-described example, the *fix* operation is useful, e.g., in cases, where the user identifies potentially interesting Complex Configurations and does not want to lose them during the continued drilldown process, but wishes to explore other Complex Configurations.

We have also considered the usefulness of the *add* operation as the same effect can be achieved by reverting filters or by using the *fix* operation. However, reverting filters can be impractical, since it can also re-add unwanted data items back to the ensemble and the desired items can also be affected by multiple filters. Fixing the items is also not a solution when the user wants to take a closer look at the Complex Configurations, but is not yet sure if they are biochemically relevant.

Our filtering operations are based on the selection of single or multiple elements. Dedicated views enable the selection of the elements at different levels of the hierarchy. As we have explained in Section 2, data items can be equipped with multiple properties, that can serve as a basis for filtering. Therefore, we have also introduced range filters that operate on the quantitative properties of the data items. Range filters can be considered as a special type of *remove complement* filters, where the items not satisfying the given range condition are removed.

Additionally, a dedicated Filter View offers the users an overview of the currently set filtering operations. Here the users can choose to revert a filter or temporarily disable some filters to see the items they affect and possibly re-add them back to the ensemble. Filters are evaluated in the order in which they were added. Exceptions to this are the *fix* operations, which override all other filters. The position of range filters in the evaluation queue is updated each time the range is changed.

Both filtering and selection operations can be performed at any level of detail. Therefore, we had to consider how to propagate the operations through the data hierarchy and the linked views. This propagation clearly needs to be bi-directional, meaning that if the operation is performed at a lower level of the hierarchy, it is first propagated to the top level and then to the lower levels again. For example, if the user removes an Amino Acid Pair (AAP), all the Complex Configurations containing this AAP will be identified and filtered from the ensemble. Consequently, all their AAPs and interacting amino acids will be removed as well. Therefore, removing one AAP in a particular view likely results in filtering other AAPs from the same view.

Similarly, if the selection propagation was automatic, the selection of one AAP in a particular view would lead to the selection of all the AAPs of all the Complex Configurations containing the initially selected AAP. However, unlike with filtering, where the removal of items upon setting a filter is expected, the automatic selection propagation is difficult to control and understand. Therefore, we decided to make the selection propagation between different levels of data hierarchy on demand.

4.2 Overview Graph

The Overview Graph (Figure 3) was designed to provide summary information about the entire ensemble (**T1**) and to aid the navigation

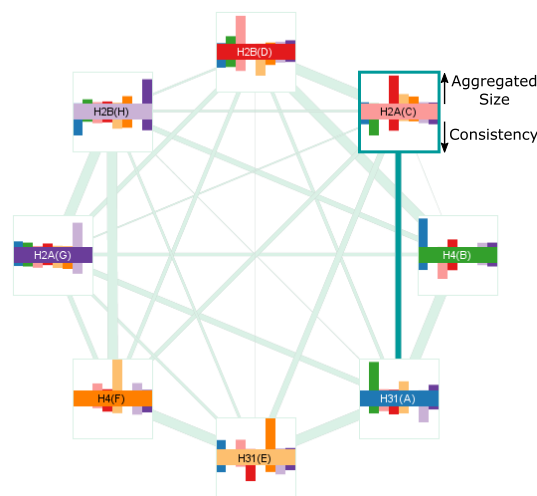


Fig. 3. Overview Graph showing a protein complex with 8 interacting proteins. The primary Protein Pair Ensemble, related to H2A(C) and H31(A), is highlighted in a darker color, as well as the primary protein H2A(C).

during the exploration process. It represents the ensemble in the form of a node-link diagram, where nodes correspond to the individual proteins of the complex, while edges between nodes represent the presence of an interaction between the corresponding proteins. We opted for a simple circular layout of the graph as it is easily readable for a small number of nodes. For protein complexes, this is sufficient as the number of proteins in a complex rarely exceeds ten.

To support the navigation to the views depicting the lower levels of the data hierarchy, which are either tied to a single protein or a pair of proteins (i.e., Protein Pair Ensemble (PPE)), the users can select a *primary protein* (node) and a *primary PPE* (edge) in the Overview Graph. These are then highlighted with a darker color (see Figure 3).

Each graph node consists of three parts. The central part of each node is colored by the color assigned to the represented protein from a predefined scheme ensuring the colors are well distinguishable. This color is then used consistently in other views to identify individual proteins. Each node is surrounded by two bar charts depicting the aggregated sizes and the consistency of the contact interfaces between the interacting proteins.

The information about the sizes of the interfaces is mapped onto the upper bar chart. The height of each bar is computed as a sum of the Amino Acid Pairs that are present in the respective Protein Pair Ensemble. The same Amino Acid Pair appearing in multiple Protein Pair Configurations is counted just once. The users can choose to either scale the bar charts independently for each protein such that the bar representing the largest interface always fills the available space or scale them using absolute values. While the former approach is better utilizing the available space, the latter allows comparing the sizes of the interfaces across the whole protein complex. However, it is less suitable for complexes where proteins vary in size – and consequently in the sizes of their contact interfaces – as the representative bars for small but still important contact interfaces may become too small.

Usually a large bar corresponds to a large contact interface between the proteins. However, since the bars represent the aggregated information, the bar for a Protein Pair Ensemble with a small but varying set of contact interfaces can be equal to the bar for a Protein Pair Ensemble with large and consistent contact interfaces. Therefore, we indicate the *consistency* of the contact interfaces in the bottom bar chart – the bigger bar signifies that the Amino Acid Pairs (AAP) in the contact interface are more consistent. The *consistency* is computed as $\frac{1}{N_{AAP}} \sum_{AAP} P_{AAP}$, where N_{AAP} is the number of unique AAPs in the contact interface (i.e., the value depicted in the upper bar chart) and P_{AAP} is the percentage of the Protein Pair Configurations (PPC) from all PPCs where the given pair is present. The light border of the bar chart provides a reference

value both for the maximum size of the interface as well as for the *consistency* equal to 1, i.e., the cases where all PPCs have exactly the same contact interface.

The width of the links between proteins encodes the number of Complex Configurations where these two proteins interact, i.e., the size of the Protein Pair Ensemble corresponding to these proteins. The users can interact with the edges via a pop-up menu to select or filter all Complex Configurations containing (or not containing) the contact between the corresponding proteins.

4.3 Property View

As we have described in Section 2, Complex Configurations (CC) can be evaluated according to various physico-chemical properties, such as the energetic score or surface complementarity. As these properties are often used for scoring and prefiltering of the configuration space, it is important for users to understand the relationship between the scoring and the dataset of Complex Configurations – i.e., the distribution of CCs according to various properties and the correlation between individual properties (**T1**, **T5**). In the Property View (see Figure 4), each CC is depicted as a point with coordinates set to two user-selected properties (i.e., we use a scatterplot representation). Users can select the CCs using rectangular brushing or clicking on individual CCs.

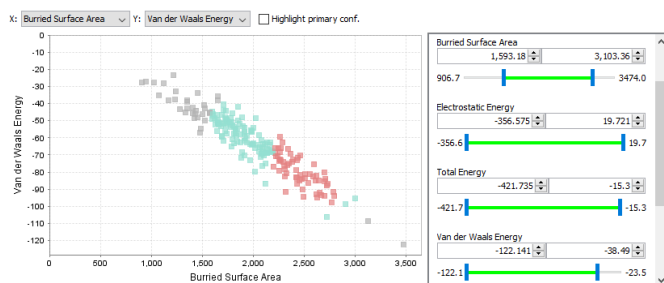


Fig. 4. Property View with the filtering panel. Selected Complex Configurations are shown in red, CCs unaffected by filters are depicted in green, and CCs filtered out by a temporarily disabled filter are shown as grey.

4.4 Protein View

In Protein View (see Figure 1d), we provide the users with more detailed information about the interaction of the selected primary protein with other proteins in the Complex Ensemble. It shows the individual amino acids (AA) of the primary protein in the first row and encodes how often these interact with other proteins. Each other row corresponds to one protein in the complex that interacts with the primary protein in at least one Complex Configuration. The cells of the row then show interactions from Protein Pair Ensemble of the primary protein and the protein corresponding to the given row. This layout enables the domain experts to identify potentially important Protein Pair Ensembles that should be further explored (**T2**). Individual rows are equipped with the name and color bar identifying individual proteins.

We map the information about the interaction frequency of AAs to the color – green for the overall number of interactions of the amino acids from the primary protein and purple for the number of interactions of these amino acids with individual proteins in other rows. The depicted frequency is always derived from the visible (i.e., unfiltered) portion of the ensemble. The coloring of the AAs of the primary protein can be changed to reflect other AA properties.

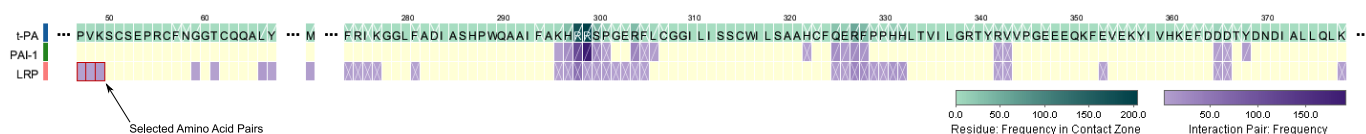


Fig. 5. Protein View of t-PA protein (primary) with the selected Amino Acid Pairs interacting between LRP protein and the Finger domain of t-PA. The AAs and Amino Acid Pairs that were filtered out based on this selection are indicated with white crosses and diagonals.

This view enables domain experts to immediately see which parts of the primary protein are interacting with other proteins (**T3**) and how often. It also provides a means to identify amino acids from the primary protein that are potentially interacting with multiple proteins. In nature, these are important but rare cases, as such bonding only happens under specific conditions (e.g., the specific spatial orientation of the amino acids). In docking predictions, it is more often the case of an incorrect bonding. The proteomic experts are, therefore, interested in locating and verifying such interactions.

The user can choose to depict either all amino acids from the primary protein or to use a condensed view (see Figure 5). In the condensed view, we hide the amino acid sequences longer than 25 amino acids that do not have any interaction. This value was experimentally chosen to remove long and uninteresting sequences from the view but also to avoid too much fracturing of the sequence, which would decrease the readability of the view as there are usually many smaller gaps between interacting amino acids. The ruler on top of the view indicates the sequence number of every 10th amino acid for improved orientation.

The view supports various filters, allowing users to filter the Complex Configurations based on the properties of individual amino acids from the primary protein (using range filters) as well as to explicitly enforce the presence (or absence) of the selected amino acids. Additionally, selecting a single row will result in setting the interface between the primary protein and protein represented by the selected row as the primary Protein Pair Ensemble that can be later explored in other views. Selection of cells will highlight the corresponding Amino Acid Pairs and AAs in other views, including the AA positions in 3D.

4.5 Residue Matrix

The Protein View allows exploring the interacting amino acids of one selected protein. However, it does not provide any detailed information about its counterparts. Therefore, to fully address task **T3**, we have adapted the solution which we designed in our previous tool [4] for the visual analysis of protein interaction pairs. When the user decides which Protein Pair Ensemble (PPE) to explore, he or she can get an aggregated overview of all Amino Acid Pairs in the selected PPE using the Residue Matrix (see Figure 1e).

Similarly to the Protein View, the axes of the Residue Matrix show the interacting amino acids from two proteins corresponding to the selected PPE. For easier navigation, the proteins are identified by the colored lines along the matrix axes. The cells in the matrix indicate the frequency of the occurrences of the Amino Acid Pair in the PPE and are also derived only from the visible portion of the ensemble. We extended the original view [4] with new options for sorting and coloring of the AAs forming the matrix axes, e.g., using properties such as charge and hydrophobicity in addition to the interaction frequency.

We further extended the original view with the support for multiple filters. The users can easily select a valid range of various biochemical properties for both individual amino acids and Amino Acid Pairs. We also allow filtering of the Complex Configuration based on the explicit manual selection of Amino Acid Pairs that need to (or cannot) be present in the PPE to be biochemically relevant.

4.6 Contact Zone List View

To address task **T5**, we have adapted another solution which we proposed in the COZOID tool [4], the Contact Zone List View (see Figure 6). This view depicts individually selected Protein Pair Configurations (PPC) side-by-side. Each PPC is represented by two columns of amino acids coming from the two proteins forming the primary Protein

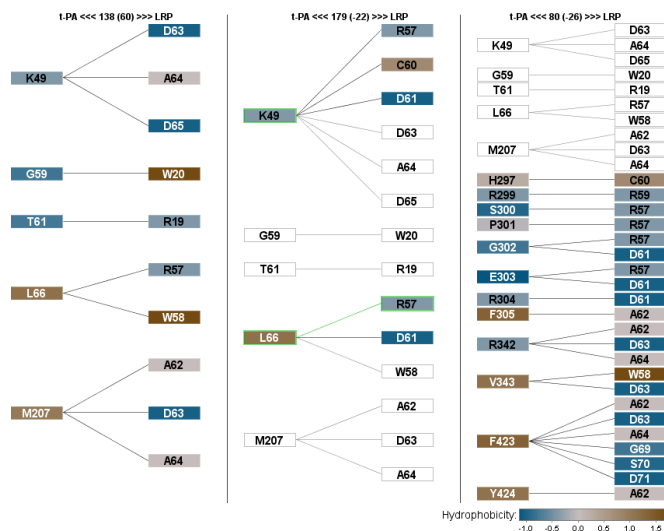


Fig. 6. Contact Zone List View showing the comparison of Amino Acid Pairs from the primary Protein Pair Configuration (PPC) (left) with two other PPCs (middle, right). AAs and Amino Acid Pairs of compared PPCs corresponding to the primary PPC are highlighted in green, while elements present in primary PPC but missing in compared PPCs are shown with empty cells. It can be seen that only the first of the compared PPCs shares some amino acids (K49, L66, R57) with the primary PPC.

Pair Ensemble. This view aims to offer a detailed representation of a small number of selected PPCs. The PPCs are by default ordered by their similarity to the reference PPC (which can be one of the PPCs in the explored ensemble or a separate PPC, e.g., of a partially resolved crystal from a protein database). The view also offers the comparison mode in which the main similarities (and differences) between the reference PPC and the rest are highlighted by the presence (and absence) of colored borders, edges, or even the whole amino acid boxes. Similarly to the Residue Matrix, we have extended this view with new options for sorting the PPCs based on the summary properties as well as sorting and coloring of the AAs in the lists.

4.7 3D Views

To depict the actual spatial arrangement of individual Complex Configurations, which is part of task **T1**, we utilize a 3D view. Our system supports all standard molecular representations which are, however, either too abstract or they suffer from occlusion and hide the most interesting parts of the protein complex – the contact interfaces between proteins. Moreover, the standard views are unusable for the exploration of larger ensembles of Complex Configurations (CCs), as showing multiple Complex Configurations at the same time results in high visual clutter and occlusion. Users can browse the individual CCs one-by-one, however, for an ensemble containing hundreds of CCs this approach is infeasible.

3D Density Overview To overcome the occlusion problem of ensembles of CCs and still giving an overview of their spatial arrangements, we propose to use an isosurface visualization (see Figure 7). First, we align individual Complex Configurations according to the selected primary protein and estimate the occurrence density of the partner proteins, i.e., all proteins in the Complex Configurations, which are interacting with the primary protein. Then, for each partner protein, we estimate the density based on its atoms, using the KDE (*kernel density estimation*). Our density estimation employs an isotropic Gaussian kernel whose bandwidth models the Van der Waals protein surface. We sample the density using a regular grid and visualize it using transparent isosurfaces showing the density levels of partner protein locations. Both the density computation and isosurface extraction are performed on the GPU, which enables efficient evaluation of the density and real-time updates of the isosurface visualization.

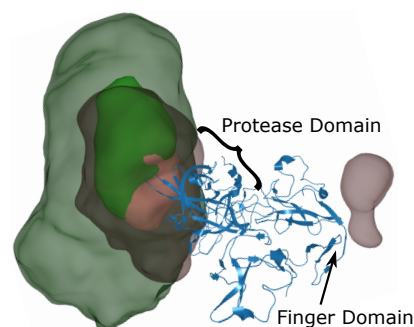


Fig. 7. 3D Density Overview of Complex Ensemble. In this case, t-PA protein (blue) was selected as the primary protein. The green and pink isosurfaces then indicate the positions of PAI-1 and LRP proteins w.r.t. the primary protein. We see that while PAI-1 seems to be interacting only with the Protease domain of t-PA, there are some configurations where LRP is interacting with the Finger domain of t-PA.

Exploded View Once the proteomic experts narrow down the set of explored Complex Configurations to units of CCs, they want to look at the 3D structure of the individual CCs and their interacting Amino Acid Pairs. To solve the problem of occlusions at the contact interfaces, we adopt the exploded view technique. This technique allows us to explode the proteins such that their relative spatial arrangement is preserved. To enable the identification of contact interfaces between the proteins, we color the atoms (and corresponding surface areas) based on the interactions between proteins. Alternatively, the coloring can be changed to reflect the properties of the contact amino acids, such as the frequency of their interaction (see Figure 1). To "explode" the Complex Configuration, we use a force-based layout for drawing graphs.

4.8 Filtering Interface

Filtering operations can be either performed based on selections of data items in the views or via dedicated range filters. Each view depicting data items with defined quantitative properties is equipped with the filtering panel that offers the range filters for the given properties. In the Property View, for example, the filtering properties correspond to the properties defined for Complex Configurations, such as electrostatics and Van der Waals energy, while in the Residue Matrix it is hydrophobicity or interaction frequency of AAs and Amino Acid Pairs. As the filtering interface can be quite spatially demanding, we only show it on demand.

We provide the users with the overview of all applied filters in a separate Filter View (see Figure 1f). It consists of two parts. The status bar at the top of the view shows how many Protein Pair Configurations were filtered out and how many are remaining in the explored ensemble. Below the status bar, the list of all applied filters is shown. For each filter, we indicate its type, subject, and properties, e.g., the type of the data item, the size of selection that was filtered, or the range of applied range filter. This way the users can see the steps they undertook during the exploration process and adjust them (**T6**). Each filter can be temporarily disabled or removed. When a filter is disabled, all filters are re-evaluated and the data items that were filtered out by this filter will reappear in the individual views with the indication that they are affected by the disabled filter. In the Overview Graph and Property View, the filtered portion of the ensemble will be grayed out. In the Protein View and Residue Matrix, we did not want too loose the color mapping that is an essential part of these views. Moreover, these views show the aggregated information, and as such, their cells may be only partially affected by the filters. This happens, for example, if a cell in the Residue Matrix represents an Amino Acid Pair present in 20 Complex Configurations, 5 of which are affected by the disabled filter. Therefore, we chose to indicate the filtered state of the items in these views by a white cross. The full cross indicates that the item is completely filtered by the disabled filter, i.e., after re-enabling of the filter the item will disappear. A single diagonal through the item cell

indicates that the cell is partially affected by the filter. This way the users can effectively study the effects of individual filters and possibly re-add the filtered items back to the ensemble. Naturally, multiple filters can be disabled at the same time and there is also the possibility to disable or remove all the filters at once.

5 INTERACTION PATTERNS

During the evaluation, when the domain experts used our tool, we observed several common interaction patterns. Since we think that these patterns are likely to occur in other cases of hierarchical data exploration, we comment on them in the following.

Upon loading an ensemble of protein complexes into our tool, users are presented with an abstract representation of the ensemble in the form of the Overview Graph. We noticed that after briefly checking this representation, and before proceeding with their analysis, the experts usually looked at a few examples from the explored ensemble using standard molecular visualization techniques. We think that they did so to improve their mental link between the new abstract representation and the actual data in the ensemble – enabling this was probably vital for the successful application of the abstraction.

We further noticed that even when the users were familiar with the explored data and knew what they were looking for, they still needed to switch multiple times between different levels of abstraction before they decided on how to proceed. We observed that users tended to spend more time focusing on the middle levels of the hierarchy (Protein View, 3D Density Overview, and Residue Matrix). This was most apparent in the early stages of the exploration when the users familiarized themselves with the dataset, as well as when searching for patterns and features that they considered for filtering. This seems to confirm that providing direct interaction, enabling the swift navigation between the individual levels of abstraction, is crucial for a successful exploration of hierarchical data.

Also the ability to relate one view to others proved to be vital. In our case, users utilized the selection operations to get more information about the selected element (e.g., selection of the Amino Acid Pairs in the Protein View, which led to their selection in the Residue Matrix, where the users could see both AAs forming the pair; and the selection of a configuration in the Property View to see its 3D structure).

We also observed a back-and-forth pattern later in the process when applying filtering operations. It became clear that disabling and reverting filters was one of the most commonly used features of our tool and that it was vital for the successful drilldown process. Therefore, providing users with explicit information about how a filtering operation affects the data proved to be crucial.

6 CASE STUDIES

Here we demonstrate the use of our tool on two case studies performed by our collaborating proteomic expert. The expert had an experience with the previous version of the COZOID tool and after a short (about 30 minutes) explanation of the new views, he was ready to use the application by himself with occasional help from the authors.

6.1 Human t-PA Model

In the first case study, the expert analyzed the results of the protein docking performed using HADDOCK multi-body docking server. The explored ensemble contained predictions of spatial arrangements for the protein complex consisting of three proteins. The first protein was tissue-type Plasminogen activator (t-PA), which is the initiator of the dissolution of blood clots in humans and other species. The second protein was Plasminogen activator inhibitor (PAI-1) that blocks the function of t-PA. It is proven to bind to the catalytic site of the Protease domain of t-PA via its reaction center loop. The third molecule was a part of the lipoprotein-like receptor (LRP) complement-type repeat. It is expected to bind either to one of the basic amino acids of the PAI-1 helices or to a Finger domain of t-PA. The input data for the docking procedure were taken from the publication of human t-PA model [24] and the crystal structure published under PDB ID 5BRR. The goal of the study was to investigate if the docking tool can predict a reasonable

configuration of the proteins that would correspond to the experimental knowledge the domain expert has about the complex.

The HADDOCK results contained 200 possible Complex Configurations of the three proteins. From these, HADDOCK selected ten Complex Configurations as the best representatives of the docking run based on the similarity clustering and internal scoring function. The proteomic expert first loaded just this subset, expecting that the most viable solution would be found in these representatives. In the Overview Graph, he selected the t-PA protein and looked at the 3D Density Overview. There he saw that in most of the cases the LRP molecule was bound close to PAI-1 and Protease domain of t-PA. However, there also seemed to be some Complex Configurations where LRP and PAI-1 were bound to the opposite side of t-PA. This observation was confirmed when the expert quickly looked on the 3D representations of individual Complex Configurations. He noticed there was precisely one Complex Configuration where this occurred. While for LRP this binding is possible, it made no sense for PAI-1 protein. Thus this Complex Configuration was removed from the ensemble. The expert also noticed that in cases where the PAI-1 was bound to the correct domain of t-PA, its orientation seemed to vary greatly. To see if the binding site, which is already well documented in the literature, was preserved despite the varying orientation, he selected the Protein Pair Ensemble of t-PA and PAI-1 for exploration in Residue Matrix. Here he selected Amino Acid Pair R299 (t-PA) and D222 (PAI-1), as these charged AAs have been proven to influence binding of the two proteins considerably, and removed the Complex Configurations that did not contain this pair. He was left with three Complex Configurations. He then switched the explored Protein Pair Ensemble to PAI-1 – LRP. After looking at the Contact Zone List View, he concluded that the position of the LRP molecule was not bound correctly in any of these cases.

Seeing that the preselected subset did not contain any viable solution, the expert then loaded a full set of 200 Complex Configurations into our system. The Overview Graph (Figure 8a) indicated that the binding sites were not very stable. Therefore, the expert again selected t-PA as the primary protein and looked at the 3D Density Overview to check the spatial distribution of the data (Figure 7). The situation was similar to the previously examined subset – both LRP and PAI-1 molecules were centered around the Protease domain of t-PA with a significant variance in positioning. The wrongly bound PAI-1 was not visible in this case (the density of such solutions was below the isosurface threshold). However, a portion of the ensemble seemed to contain LRP binding to the Finger domain of t-PA.

Based on the experience with the subset of Complex Configurations, the expert expected that the large portion of t-PA – PAI-1 Protein Pair Ensemble would be infeasible. Therefore, he selected this Protein Pair Ensemble and repeated the Amino Acid Pair filtering in the Residue Matrix (R299 (t-PA) – D222 (PAI-1)). This operation immediately removed all but 35 Complex Configurations.

The expert wanted to see how the remaining Complex Configura-

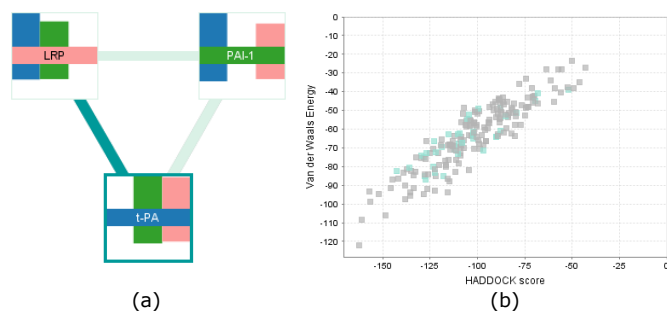


Fig. 8. a) Overview Graph of the tPA – PAI-1 – LRP docking results. The small height of the colored bars below the core of each node indicates that the contact interfaces between the proteins are not very stable. b) Property View showing the relationship between the HADDOCK score and Van der Waals energy. The filtered out Complex Configurations are depicted in gray.

tions performed according to different scoring functions, curious why the HADDOCK selected representatives did not contain more biochemically relevant solutions. He wanted to compare the scores of already filtered out Complex Configurations to the ones he was still considering. Therefore, he temporarily disabled the filtering and turned his attention to the Property View. As he browsed through different combinations of properties, he observed that the HADDOCK score was closely related to Van der Waals energy (Figure 8b). However, none of the properties seemed to correspond to his own filtering in any way. So he concluded that in this instance he could not rely on the pre-computed scores.

The expert then decided to proceed with the exploration of the LRP bindings. He knew from the 3D Density Overview that in the majority of Complex Configurations the LRP molecule was located close to the t-PA – PAI-1 binding site and decided to explore this part of the dataset first. Thus he wanted to remove the Complex Configurations where this was not the case. As the t-PA was still selected as the primary protein, the expert switched to the Protein View to locate the Finger domain amino acids with the LRP binding. He selected the corresponding Amino Acid Pairs and Complex Configurations where these Amino Acid Pairs were interacting (Figure 5) and then filtered out the selection. This operation removed only three Complex Configurations.

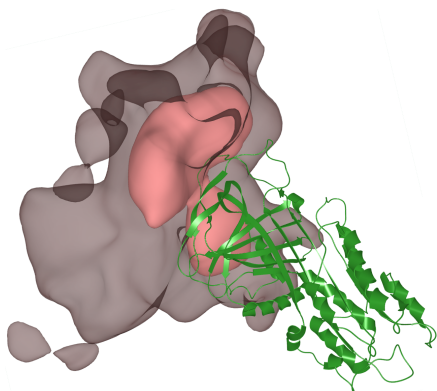


Fig. 9. 3D Density Overview of the subset of 35 Protein Pair Configurations of PAI-1 (green) with LRP (pink).

The expert knew that LRP was proven to bind to a basic amino acid on the PAI-1 helix. Therefore, he wanted to find out if any of the Complex Configurations fulfilled this condition. He switched the primary protein to PAI-1. In the 3D Density Overview he saw that LRP molecules in the ensemble were close enough to only one helix of PAI-1 – the helix closest to the binding site with t-PA (Figure 9). In the Protein View, he looked up the amino acid sequence part corresponding to this helix and its only basic amino acid, H261. He removed the Complex Configurations where LRP had no interaction with this amino acid and ended up with only one Complex Configuration. However, looking at its 3D structure (Figure 10), he was immediately able to see that this Complex Configuration was improbable since the location of LRP was acting as a wedge between t-PA and PAI-1.

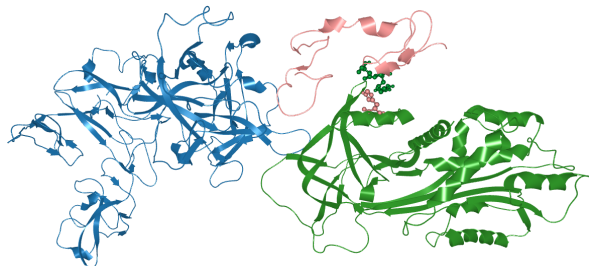


Fig. 10. Complex Configuration with LRP (pink) interacting with basic residue H261 on PAI-1 (green) helix. It can be seen that the position of LRP disrupts the interaction between t-PA (blue) and PAI-1.

Seeing that he was unable to find a relevant Complex Configuration with LRP binding to the PAI-1, the expert decided to explore the smaller subset where LRP was binding to the Finger domain of t-PA. He removed all the filters except for the first one, returning to the point where he had 35 Complex Configurations. He then repeated the selection of Complex Configurations with interactions at the Finger domain of t-PA, this time removing the complementary subset, so he ended up with three Complex Configurations. Using the Contact Zone List View and 3D representations of individual Complex Configurations, he selected the one with the tightest contact with t-PA (Figure 11). The expert claimed that this Complex Configuration was not expected but it has possible biological importance, which should be further studied.

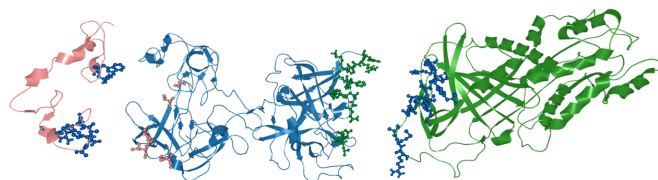


Fig. 11. Exploded view of the Complex Configuration with the best binding between t-PA (blue) and LRP (pink) showing the interacting AAs. It can be seen that unlike the case in Figure 10, the interaction between the t-PA and PAI-1 is preserved here.

He further wanted to find out if there were more similar Complex Configurations with regard only to the binding at the Finger domain of t-PA (i.e., disregarding the t-PA – PAI-1 interaction). He therefore temporarily disabled the filters and again selected all Complex Configurations with interaction at the Finger domain. This operation also selected previously filtered Complex Configurations. The expert then re-added these Complex Configurations to the ensemble, which yielded 41 Complex Configurations. He then sorted the Property View according to the similarity to the initially selected Complex Configuration. He saw that most of the Complex Configurations exhibited quite a low similarity. He selected five Complex Configurations with the highest scores to compare in the Contact Zone List View and 3D, but concluded that the first structure was still the best among them. The Contact Zone List View showing the t-PA – LRP Protein Pair Configurations of the best Complex Configuration and the two most similar Complex Configurations is shown in Figure 6.

In this case study, the proteomic expert analyzed an ensemble of 200 Complex Configurations. After he checked the Complex Configurations preselected by the docking tool using his usual tools, he stated that he does not believe the docking tool was successful in predicting a potentially relevant Complex Configuration. We then asked the domain expert to re-evaluate the ensemble using our solution. He again started with the analysis of the preselected Complex Configurations, which confirmed his initial statement. However, he then continued with the analysis of the full ensemble, where he eventually found one Complex Configuration that he considered possibly biologically important and that should be explored further. He concluded that finding this Complex Configuration without our solution would be virtually impossible. He would not be inclined to explore this dataset in detail, because it would take him an incredible amount of time and given the initial results, he did not think there was a chance of finding anything interesting. He stated that even if he chose to explore the ensemble in greater detail, he would easily miss the one Complex Configuration he selected in the end, as it did not have the expected spatial arrangement. However, with our solution, he was easily able to filter out unsuitable Complex Configurations, focus on interesting subsets, and find their best representatives. Our solution enabled him to explore the ensemble in a faster manner and discover new and unexpected results.

6.2 Human Nucleosome

To verify the scalability of our solution, we wanted to test it also on an ensemble of a larger protein complex. Therefore, we created an ensemble of prediction for human nucleosome structure (published

under PDB ID 3AFA). This structure consists of eight proteins. The ensemble was prepared by taking pairwise predictions for five pairs of proteins produced by pyDock [2], which were then iteratively combined to recreate the nucleosome structure. In this phase, the colliding combinations were automatically discarded. Then, a representative sample of 500 solutions was selected, which included a set of 35 Complex Configurations closest to the published crystal structure as well as randomly sampled decoys.

We then loaded the data into our system and asked the proteomic expert who provided us with the data to identify the best solutions. The expert was already familiar with the interaction patterns in the nucleosome structure and thus he began his exploration in the Overview Graph, where he removed Protein Pair Ensembles which were not supposed to interact. He then continued his analysis using the Residue Matrix, where he continued with the filtering based on the most prominent contacts (e.g., the salt bridge between units H2A and H2B), until he finished with a set of 11 configurations. He then loaded the crystal structure to our tool and used the Contact Zone List View to compare it with the selected structures. He found out that the majority of structures was reasonably similar to the crystal structure. A snapshot from this exploration can be seen in Figure 1. In terms of scalability, the tool was fully interactive even for this larger ensemble. The performance of the tool only differs in the initial precomputation time, which may take several minutes, depending on the size of the ensemble. However, the precomputation time is not a limitation for the experts.

7 FEEDBACK AND DISCUSSION

We have evaluated our system with the domain experts from two different research groups. Here we summarize the most notable feedback and also discuss some limitations of our approach.

The proteomic experts stated that the Overview Graph together with the 3D Density Overview provide a nice initial overview of the ensemble. They particularly liked the graph stating that it is useful especially for protein complexes consisting of a higher number of protein units, as the interaction patterns become more intricate and harder to analyze with the increasing size of the complex.

When designing our solution, we have considered multiple layouts for the Overview Graph. In the end, we decided to stick to the circular layout as it communicates that the visualization is abstract. The non-uniform layout could suggest that some proteins are spatially closer than the others which would be misleading as we cannot satisfactorily compute these spatial properties from the dataset. We have also considered different mappings (instead of the bar chart) to communicate the size of the contact interfaces. However, as the contact interface size is tightly connected with its consistency, we always need to depict these two values close to each other. If we would choose, for instance, the edges of the graph for this mapping, we would also need two information channels, i.e., edge width and its color. We do not consider this a suitable combination, as the color interpretation might be distorted due to the varying width of the lines.

Regarding the filtering interface, the users appreciated the ability to disable the filters and observe their effect as well as the possibility to revert filters and thus return to the previous points in exploration. They noted that it would be nice to have the possibility to group the applied filters and annotate them with their own notes so it would be easier to navigate in the filter list and they could disable or remove the groups together. We plan to add the support for filter grouping in the near future. However, considering that the results of filtering depend on the order of filters (in particular the position of the *add* operation in the filter queue), the grouping needs to be restricted to preserve the initial order of the filters.

Related to this was the request to save the exploration state for future analysis. Our framework already partially supports this via the ability to store workspaces, which includes storing of selected configurations, and we are currently extending it to support also the storage of filter states.

The proteomic experts also listed several properties that could be added to our tool, such as amino acid conservation or the type of bond in the Amino Acid Pairs. Our solution can be easily extended with

these. On a similar note, we would also like to provide the users with the support for importing the precomputed Complex Configuration scores from additional docking tools. According to domain experts, it is more reliable and informative to use the scores computed during docking simulation than re-scoring them afterwards.

Overall, the proteomic experts appreciated the ability to navigate through different levels of data hierarchy and use them to filter and select the Complex Configurations. They stated that our tool enabled them to explore their ensemble in a manner that was not previously possible. It considerably decreased the time spent on the analysis process and enabled them to identify solutions they would otherwise not be able to find.

We would also like to discuss the limitations and scalability issues of our application. The Overview Graph is limited by the number of proteins in a studied protein complex. As mentioned, this number in known complexes rarely exceeds 10. However, we cannot assure that this number will not be higher in the future. In such cases, the Overview Graph may become too cluttered to be used effectively. It is also limited by the number of colors that are currently used to distinguish between individual proteins, as it is not recommended to use more than 7–12 colors for such an application [5].

Regarding the performance scalability, the most demanding step is the preprocessing of the data, which includes the computation of the contact interfaces, various properties, and initialization of the hierarchical data structures. The performance in this step has an asymptotic complexity $\mathcal{O}(m^2 * n)$ where m is the number of proteins in the complex and n is the number of individual Complex Configurations. We have tested the application on the dataset containing 500 Complex Configurations of a protein complex consisting of 8 proteins. Here the initialization step took several minutes. However, after this the application was fully interactive in real time.

8 CONCLUSION AND FUTURE WORK

We have presented a novel approach to the analysis of large ensembles of multi-body protein complexes. This approach solves the problem of analyzing hundreds of possible spatial configurations of protein complexes produced by a docking tool and selecting suitable representatives that can be further explored in the lab. Our approach builds on the hierarchical structure of protein complex ensembles which we use to facilitate the exploration of the data at multiple levels of the hierarchy.

We have evaluated the usefulness of our approach in a case study performed by a proteomic expert. In the case study, the expert was able to use our tool and find an important configuration of the explored protein complex in the ensemble, which he initially marked as uninteresting based on preselected samples. This was a previously unattainable task, as the domain experts lacked proper tools to explore the entire ensemble in detail. In the future, we would like to extend our solution with the possibility to automatically cluster the results based on the similarity of the contact interfaces and spatial arrangement, and provide means for comparison of groups of Complex Configurations.

Another topic that came up during the design process of our system was to search for a suitable scoring function. As we have stated in Section 2, different tools use different scoring functions that are based on a combination of energetic, physical, and evolutionary properties and there is no single best option. Thus finding a suitable scoring function is a challenging task dealing with multi-objective optimization and it could benefit greatly from our visual guidance.

ACKNOWLEDGMENTS

We would like to thank the domain experts Jan Mičan and Jan Paleček for providing us with the datasets, valuable feedback, and conducting the case study. This work was supported by the Czech Science Foundation international project GC18-18647J and the Internal Masaryk University grant (MU/0822/2015). Parts of this work were done in the context of the VIDÍ project, which is supported by Bergens Forskningsstiftelse, the Mohn Medical Imaging and Visualization Center (MMIV), and the University of Bergen.

REFERENCES

- [1] N. Amir, D. Cohen, and H. J. Wolfson. DockStar: a novel ILP-based integrative method for structural modeling of multimolecular protein complexes. *Bioinformatics*, 31(17):2801–2807, 2015. doi: 10.1093/bioinformatics/btv270
- [2] T. M.-K. Cheng, T. L. Blundell, and J. Fernandez-Recio. pyDock: electrostatics and desolvation for effective scoring of rigid-body protein–protein docking. *Proteins: Structure, Function, and Bioinformatics*, 68(2):503–515, 2007. doi: 10.1002/prot.21419
- [3] J. Esquivel-Rodríguez, Y. D. Yang, and D. Kihara. Multi-LZerD: Multiple protein docking for asymmetric complexes. *Proteins: Structure, Function, and Bioinformatics*, 80(7):1818–1833, 2012. doi: 10.1002/prot.24079
- [4] K. Furmanová, J. Byška, E. M. Gröller, I. Viola, J. J. Paleček, and B. Kozlíková. COZOID: contact zone identifier for visual analysis of protein–protein interactions. *BMC Bioinformatics*, 19(1):125, 2018. doi: 10.1186/s12859-018-2113-6
- [5] M. Harrower and C. A. Brewer. Colorbrewer.org: an online tool for selecting colour schemes for maps. *The Cartographic Journal*, 40(1):27–37, 2003.
- [6] S.-Y. Huang. Search strategies and evaluation in protein–protein docking: principles, advances and challenges. *Drug Discovery Today*, 19(8):1081–1096, 2014. doi: 10.1016/j.drudis.2014.02.005
- [7] S.-Y. Huang and X. Zou. An iterative knowledge-based scoring function for protein–protein recognition. *Proteins: Structure, Function, and Bioinformatics*, 72(2):557–579, 2008. doi: 10.1002/prot.21949
- [8] W. Humphrey, A. Dalke, and K. Schulten. VMD: Visual molecular dynamics. *Journal of Molecular Graphics*, 14(1):33 – 38, 1996. doi: 10.1016/0263-7855(96)00018-5
- [9] Y. Inbar, H. Benyamini, R. Nussinov, and H. J. Wolfson. Prediction of multimolecular assemblies by multiple docking. *Journal of Molecular Biology*, 349(2):435–447, 2005. doi: 10.1016/j.jmb.2005.03.039
- [10] J. Janin, K. Henrick, J. Moult, L. T. Eyck, M. J. Sternberg, S. Vajda, I. Vakser, and S. J. Wodak. Capri: a critical assessment of predicted interactions. *Proteins: Structure, Function, and Bioinformatics*, 52(1):2–9, 2003.
- [11] A. Jurčík, D. Bednář, J. Byška, S. M. Marques, K. Furmanová, L. Daniel, P. Kokkonen, J. Brezovský, O. Strnad, J. Štourač, A. Pavelka, M. Maňák, J. Damborský, and B. Kozlíková. CAVER Analyst 2.0: analysis and visualization of channels and tunnels in protein structures and molecular dynamics trajectories. *Bioinformatics*, 34(20):3586–3588, 2018. doi: 10.1093/bioinformatics/bty386
- [12] E. Karaca, A. S. Melquiond, S. J. de Vries, P. L. Kastiris, and A. M. Bonvin. Building macromolecular assemblies by information-driven docking introducing the HADDOCK multibody docking server. *Molecular & Cellular Proteomics*, 9(8):1784–1794, 2010. doi: 10.1074/mcp.M000051-MCP201
- [13] G. Kuzu, O. Keskin, R. Nussinov, and A. Gursoy. PRISM-EM: template interface-based modelling of multi-protein complexes guided by cryo-electron microscopy density maps. *Acta Crystallographica Section D: Structural Biology*, 72(10):1137–1148, 2016. doi: 10.1107/S2059798317017739
- [14] R. A. Laskowski, E. G. Hutchinson, A. D. Michie, A. C. Wallace, M. L. Jones, and J. M. Thornton. PDBsum: a Web-based database of summaries and analyses of all PDB structures. *Trends in Biochemical Sciences*, 22(12):488–490, 1997. doi: 10.1016/S0968-0004(97)01140-7
- [15] R. A. Laskowski and M. B. Swindells. Ligplot+: Multiple ligand/protein interaction diagrams for drug discovery. *Journal of Chemical Information and Modeling*, 51(10):2778–2786, 2011. doi: 10.1021/ci200227u
- [16] C. H. Lee and A. Varshney. Computing and displaying intermolecular negative volume for docking. In G.-P. Bonneau, T. Ertl, and G. M. Nielson, eds., *Scientific Visualization: The Visual Extraction of Knowledge from Data*, pp. 49–64. Springer Berlin Heidelberg, Berlin, Heidelberg, 2006. doi: 10.1007/3-540-30790-7_4
- [17] E. D. Levy, J. B. Pereira-Leal, C. Chothia, and S. A. Teichmann. 3D complex: a structural classification of protein complexes. *PLoS Computational Biology*, 2(11):e155, 2006. doi: 10.1371/journal.pcbi.0020155
- [18] T. Lu, J. H. Ding, and S. N. Crivelli. DockingShop: a tool for interactive protein docking. In *2005 IEEE Computational Systems Bioinformatics Conference - Workshops (CSBW'05)*, pp. 271–272, 2005. doi: 10.1109/CSBW.2005.54
- [19] H. Miao, T. Klein, D. Kouřil, P. Mindek, K. Schatz, M. E. Gröller, B. Kozlíková, T. Isenberg, and I. Viola. Multiscale molecular visualization. *Journal of Molecular Biology*, 431(6):1049 – 1070, 2019. doi: 10.1016/j.jmb.2018.09.004
- [20] S. I. O’Donoghue, K. S. Sabir, M. Kalemánov, C. Stolte, B. Wellmann, V. Ho, M. Roos, N. Perdigo, F. A. Buske, J. Heinrich, B. Rost, and A. Schafferhans. Aquaria: simplifying discovery and insight from protein structures. *Nature Methods*, 12(2):98–99, 2015. doi: 10.1038/nmeth.3258
- [21] J. Paleček. SMC5/6: Multifunctional player in replication. *Genes*, 10(1):7, 2019. doi: 10.3390/genes10010007
- [22] E. F. Pettersen, T. D. Goddard, C. C. Huang, G. S. Couch, D. M. Greenblatt, E. C. Meng, and T. E. Ferrin. UCSF Chimera—a visualization system for exploratory research and analysis. *Journal of Computational Chemistry*, 25(13):1605–1612, 2004. doi: 10.1002/jcc.20084
- [23] M. Pikora and A. Gieldon. RASMOL AB - new functionalities in the program for structure analysis. *Acta Biochimica Polonica*, 62(3):629–631, 2015. doi: 10.18388/abp.2015.972
- [24] Y. S. Rathore, M. Rehan, K. Pandey, and G. Sahní. First structural model of full-length human tissue-plasminogen activator: a SAXS data-based modeling study. *The Journal of Physical Chemistry B*, 116(1):496–502, 2011. doi: 10.1021/jp207243n
- [25] S. Reille, M. Garnier, X. Robert, P. Gouet, J. Martin, and G. Launay. Identification and visualization of protein binding regions with the ArDock server. *Nucleic Acids Research*, 46(W1):W417–W422, 2018. doi: 10.1093/nar/gky472
- [26] R. A. Sayle and E. J. Milner-White. RASMOL: biomolecular graphics for all. *Trends in Biochemical Sciences*, 20(9):374, 1995. doi: 10.1016/S0968-0004(00)89080-5
- [27] Schrödinger, LLC. The PyMOL molecular graphics system, version 1.8. November 2015.
- [28] P. Šledz and A. Cafilisch. Protein structure-based drug design: from docking to molecular dynamics. *Current Opinion in Structural Biology*, 48:93–102, 2018. doi: 10.1016/j.sbi.2017.10.010
- [29] R. Splechtna, K. Matković, D. Gračanin, M. Jelović, and H. Hauser. Interactive visual steering of hierarchical simulation ensembles. In *2015 IEEE Conference on Visual Analytics Science and Technology (VAST)*, pp. 89–96, 2015. doi: 10.1109/VAST.2015.7347635
- [30] A. Vangone, R. Oliva, and L. Cavallo. CONS-COCOMAPS: a novel tool to measure and visualize the conservation of inter-residue contacts in multiple docking solutions. *BMC Bioinformatics*, 13 Suppl 4:S19, 2012. doi: 10.1186/1471-2105-13-S4-S19
- [31] A. Vangone, R. Spinelli, V. Scarano, L. Cavallo, and R. Oliva. COCOMAPS: a web application to analyze and visualize contacts at the interface of biomolecular complexes. *Bioinformatics*, 27(20):2915–2916, 2011. doi: 10.1093/bioinformatics/btr484
- [32] V. Venkatraman and D. W. Ritchie. Predicting multi-component protein assemblies using an ant colony approach. *International Journal of Swarm Intelligence Research (IJSIR)*, 3(3):19–31, 2012. doi: 10.4018/ijrir.2012070102