le cnam

Comparative Evaluation of Protein-Protein Docking Algorithms in SHREC2024

Master's Thesis in Molecular Bioinformatics: Methods and Analyses

Supervised by:

Prof. Matthieu MONTES

Dr. Florent LANGENFELD

Dr. Taher YACOUB

Academic Year 2023-2024

by Nika ZARUBINA

Acknowledgements

I am deeply grateful to my supervisor, Matthieu, for providing me with this incredible opportunity to learn and grow. His guidance and support through every challenging deadline have been invaluable. Throughout this journey, I have gained so much knowledge and experience.

My sincere thanks to Florent, whose willingness to answer all my questions with patience and kindness made a significant difference. Thank you to Taher, an extraordinary person, tutor, and mentor, whose wisdom and gentle encouragement have always guided me in the right direction without ever imposing.

I would also like to express my heartfelt thanks to Nathalie for her kind approach and her humorous and gentle support, which has been a source of comfort and motivation.

I am grateful to the Gradient project and its dedicated members for giving me the opportunity to be a part of this meaningful internship. Finally, I want to extend my deepest appreciation to my fellow interns, Mariyam, Matilde, and Daniela, Henes, Baptiste, Samuel, Alice for the wonderful friendships and support we shared. This experience has been enriched by their presence and camaraderie.

Abstract

The SHREC2024 track aims to compare different algorithms for retrieving non-rigid complementary shape pairs, specifically applied to complex 3D objects such as proteins. This benchmark utilized a dataset of 52 selected protein-protein complexes with publicly available experimental structures. A key challenge of this task is the exclusion of shapes derived from ground truth conformations from the dataset. Various metrics were employed to evaluate retrieval performance, including nearest-neighbor, first-tier, second-tier, and true positives, as well as the quality of predicted poses using TM-score, lDDT, ICS, IPS, and DockQ—metrics commonly used in the Critical Assessment of PRediction of Interactions (CAPRI) challenges. Two teams participated in this challenge and successfully returned the expected results. This paper discusses these results and explores the prospects of retrieval methods based solely on protein shape information in the absence of atomic data, within the broader context of protein-protein docking.

Table of Contents

Table of Softwares

Python	3.12.0
EDTSurf	0.2009
OpenStructure	2.7
RANSAC	
Meshlab	2023.12

Table of Abbreviations

Abbreviation	**Meaning**
CAPRI	Critical Assessment of Predicted Interaction
FT	First-Tier
ICS	Interface Contact Similarity)
IPS	Interface Patch Similarity)
LDDT	Local Distance Difference Test
NN	Nearest-Neighbor
PPIs	Protein-protein interactions
RANSAC	Random Sample Consensus
SES	Solvent Excluded Surface
SHREC	Shape Retrieval Challenge
ST	Second-Tier
TM-score	Template Modeling score
TP	True Positive

List of Figures

Introduction

1.1 Proteins: structure and function

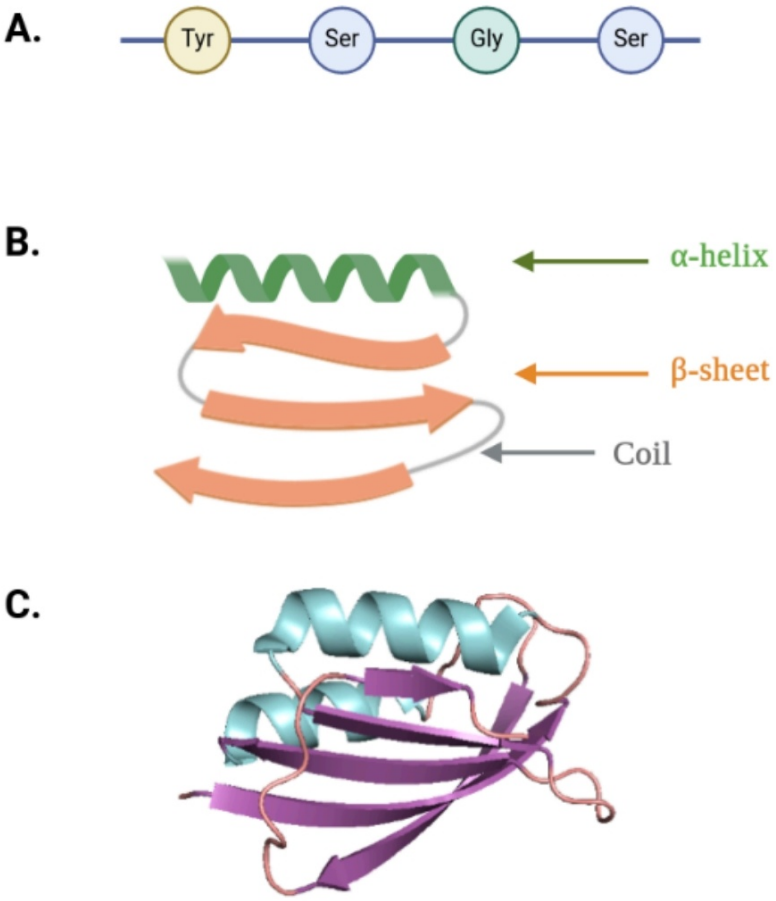

Figure 1: Representation of protein structures: A) Primary structure; B) Secondary structure with alpha-helices in green and beta-sheets in orange; C) Tertiary structure of crystal structure of acylphosphotase (code PDB: 1W2I)

Proteins are large, complex molecules composed of long chains of amino acids. These amino acids are linked together by covalent peptide bonds to form a polypeptide chain. The specific sequence of amino acids in a polypeptide chain determines the protein's structure and function, known as the primary structure of the protein. The polypeptide chain folds into specific three-dimensional shapes essential for the protein's activity. Proteins can be represented at multiple levels, each providing different insights into their struc-

ture and function. The primary structure represents the linear sequence of amino acids, giving a concise but limited view of the protein. The local folding of the polypeptide chain into structures such as α-helices and β-sheets constitutes the protein's secondary structure. α-helices are spiral-shaped and stabilized by hydrogen bonds, while β-sheets are formed by strands connected laterally by hydrogen bonds, creating a sheet-like array. The tertiary structure provides a comprehensive 3D view of a single polypeptide chain, revealing how secondary structures fold together (Figure 1).

This 3D structure is crucial for the protein's biological activity as it determines the spatial arrangement of functional groups and active sites.

1.2 Relationship between sequence/structure/surface

The **sequence**, or primary structure, is the linear order of amino acids joined by peptide bonds in a polypeptide chain. This sequence is fundamental because it dictates the protein's folding and eventual function. The **secondary structure** refers to local conformations within the polypeptide chain, such as α-helices and β-sheets, which are stabilized by hydrogen bonds between the backbone atoms. The **tertiary structure** encompasses the overall three-dimensional shape of a single polypeptide chain, formed by the interactions among the side chains of the amino acids, including hydrogen bonds, ionic interactions, hydrophobic packing, and disulfide bridges.

The **surface** of a protein is a higher-level representation that captures the accessible regions of the tertiary structure. Unlike the sequence and internal structure, which focus on the linear and spatial arrangement of amino acids, the surface emphasizes the external features that interact directly with other molecules, such as substrates, inhibitors, or other proteins. These interactions are mediated by the chemical properties of the surface residues, including charge, hydrophobicity, and the presence of functional groups. The surface is crucial for understanding protein function because it is where binding and recognition events occur, which are essential for processes like enzyme activity, signal transduction, and cellular communication.

1.3 Surface of proteins

There are three primary types of protein surfaces that are commonly analyzed (Figure 2):

1. **Van der Waals Surface** [4]: The van der Waals surface is based on the van der Waals radii, which approximates the effective size of an atom as a sphere. This surface represents the boundary formed by the union of these spheres, each defined by the van der Waals radius of its constituent atoms. The van der Waals surface provides a basic model of the protein's outer boundary, reflecting the physical space occupied by the molecule. Importantly, these spheres are impenetrable, meaning that no other object can penetrate this surface, as it represents the hard physical boundary of the protein.

2. **Solvent-Accessible Surface (SAS)** [17]: The solvent-accessible surface is calculated by rolling a spherical probe, typically the size of a water molecule (1.4 angstroms), over the van der Waals surface of the protein. The path traced by the center of the probe defines the SAS, indicating the regions of the protein that are accessible to solvent molecules.

3. **Solvent-Excluded Surface (SES)** [8]: Also known as the Connolly surface, the solvent-excluded surface is derived from the SAS. It is defined by the points of contact between the probe and the van der Waals surface. When the probe touches the van der Waals surface at two points, the surface is traced along the path connecting these contact points. This method captures the contours of narrow cavities and grooves on the protein surface, offering a more detailed and realistic representation of the protein's surface topology.

1.4 Protein-Protein Interactions

Protein activities are primarily mediated through interactions between proteins. These interactions are driven by the complementarity of their molecular surfaces, where the shapes and chemical properties of protein surfaces match and fit together to form complex assemblies. Understanding these interactions is crucial for comprehending the intricate workings of life. Protein-protein interactions (PPIs) are at the core of almost all biological processes, and disruptions in these interactions can lead to various diseases, including cancer, neurodegenerative disorders, and immune deficiencies [9]. By studying PPIs, scientists can gain insights into cellular mechanisms and develop targeted therapies to treat diseases linked to dysfunctional protein interactions.

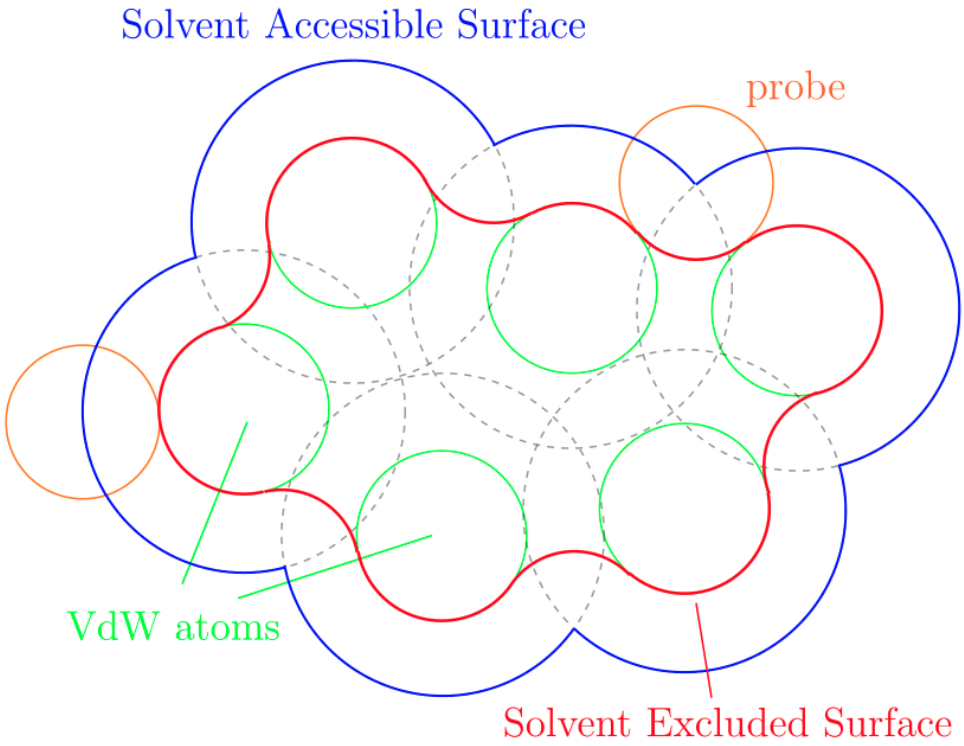

Figure 2: A 2D schematic representation illustrating the Van der Waals surface (in green), the Solvent Accessible Surface (in blue), and the Solvent Excluded Surface. Extract from Chaoyu Quan et al.(2018) [20]

1.5 Critical Assessment of Predicted Interactions (CAPRI) challenge

In the field of structural biology, numerous methods are dedicated to identifying structural similarities in proteins. These techniques often rely on the analysis of three-dimensional structures, which are defined by the atomic coordinates of the protein's constituents.

CAPRI is a community-wide experiment aimed at assessing the accuracy of methods for predicting protein-protein interactions. CAPRI was established to promote the development and evaluation of computational techniques that predict the structures of protein-protein complexes [14]. CAPRI aims to evaluate the performance of computational algorithms in predicting the structures of protein-protein interactions. This involves predicting how two proteins will dock together to form a complex based on their individual structures.

In CAPRI experiments, organizers provide participants with the 3D structures of individual proteins (referred to as "targets") that are known to form a complex. Participants then use their computational methods to predict the structure of the complex. These

predictions are compared against the experimentally determined structures (often obtained through methods like X-ray crystallography or cryo-electron microscopy), which are revealed only after the predictions have been submitted. Predictions are assessed based on various criteria, including the accuracy of the predicted interface (where the two proteins interact), the overall structural alignment with the experimental complex, and the ability to correctly identify the interacting regions.

1.6 Difficulties of CAPRI challenge

Traditionally, researchers have focused on analyzing the three-dimensional structures of proteins to understand their functions and interactions. However, it's important to recognize that differences in protein structure do not always translate to differences in surface properties. Molecular surfaces represent an abstraction of the protein's underlying structure, encapsulating geometric and chemical features essential for interactions. Despite variations in sequence and backbone conformation, proteins often share similar surface properties [18].

By examining these 3D structures, researchers can determine surface complementarity, which is crucial for understanding PPIs. Surface complementarity refers to the way in which the shapes and chemical properties of protein surfaces match and fit together, allowing proteins to align and bind effectively to form functional complexes. This principle is fundamental in processes such as enzyme-substrate binding, antibody-antigen recognition, and the formation of multi-protein assemblies. To address these challenges, methods like protein-protein docking have been developed to predict protein interactions based on surface shape and complementarity [12].

1.7 SHREC2024:

Community benchmarks such as the Shape Retrieval Challenge (SHREC) play a crucial role in evaluating the performance of algorithms in retrieving and comparing shapes in different domains. SHREC is an annual competition in the field of computer vision and image processing. It aims to evaluate the performance of algorithms for retrieving and comparing three-dimensional shapes. Participants are tasked with developing methods to retrieve similar shapes from large databases of 3D models, often focusing on specific

applications such as object retrieval in point clouds, anatomical shape retrieval in medical data ([10], [16]).

This year there are two tracks: Recognition of hand motions molding clay and Protein shapes docking which concerns us. The objective of this track is to evaluate the effectiveness of shape complementarity algorithms using a dataset containing protein-protein interactions.

1.7.1 Difficulty of the protein shapes docking track

Finding compatible protein pairs for complex formation is a formidable task in structural biology, as highlighted by the results of challenges like CAPRI. Typically, this challenge is approached using structure-based methods that analyze the arrangement of atoms in proteins. However, this year's SHREC track proposes to focus on surface characteristics to address this challenge. It's important to note that in this context, we are solely considering surface features and not the overall structure of the proteins, emphasizing the unique approach taken in the SHREC challenge. Unlike traditional methods, surface-based approaches consider the geometric and chemical features of protein surfaces, offering a more comprehensive perspective on PPIs.

Proteins are dynamic biomolecules whose structures can vary, especially upon binding. This variability presents an additional hurdle in predicting how protein shapes interact, particularly when their unbound structures differ from their bound counterparts. Moreover, intra-class variability, characterized by different deformations of the same protein due to non-rigid behavior, compounds this challenge (Figure 3). This variability is a consequence of the flexibility of protein chains, influenced by the internal degrees of freedom in their rotational bonds.

It's interesting to know that currently, there is no available dataset containing both bound complexes and unbound protein shapes. Through this challenge, we hope to not only benchmark current methodologies but also to inspire the development of innovative approaches that can more accurately and efficiently predict protein interactions, ultimately contributing to advancements in computational biology and bioinformatics.

Figure 3: A superposition of the unbound Elongation Factor 2 (EF2) protein (blue) and the EF2 protein (yellow) bound to exotoxin A. Exotoxin A is not shown in this representation.)

1.8 Objectives of our participation in SHREC2024

In this SHREC2024 track, we address the challenge of 3D object complementarity applied to protein-protein complexes with two goals:

- the ability to retrieve the two actual chains of each of the 52 experimentally-resolved complexes in the benchmark
- the ability to identify the true interface area between them.

1.9 Constraints and objectives for participants

1.9.1 Constraints

1. **Timeline Constraints**: Participants were given 7 weeks from the dataset's release to submit their results. Several teams expressed interest in the track, but only two managed to return the required results within the time limit.
2. **Lack of Publicly Available Reference Datasets**: There is a lack of publicly available datasets of protein complexes in the form of 3D meshes or point clouds. 3D meshes are digital representations of protein surfaces composed of interconnected vertices, edges, and faces, creating a continuous surface. On the other hand, point

clouds are discrete representations of protein surfaces, where each point represents a sampling point on the surface. Such datasets are crucial for training machine-learning algorithms designed for predicting protein-protein interactions. The absence of these resources posed a significant challenge for participants. This track's dataset may represent a crucial first step toward the creation and publication of curated datasets, which are essential for developing and training new surface-based methods for predicting molecular complexes.

1.9.2 Participants' objectives

Participants were supplied with two sets of protein shapes, categorized as queries and targets. They were tasked with two primary objectives:

1. Evaluating Likelihood: Participants had to assess the likelihood of each query-target complex. This involved analyzing the compatibility and potential interactions between the protein surfaces.

2. Predicting Complexes: Participants were required to provide the ten most probable predicted complexes for each query-target pair. This step was critical for determining the effectiveness and accuracy of their methods in predicting realistic protein-protein interactions.

By addressing these challenges and working with the provided datasets, the participants contributed to advancing the field of protein docking.This initiative highlights the importance of collaborative efforts and shared resources in order to tackle complex scientific problems.

Methods

2.1 Participants

The participants of this track are:

- **Kihara et *al.***: They employed LZerD, a surface-based protein docking method developed by their group. LZerD works by exhaustively sampling possible interaction interfaces and angles between two protein structures. It uses 3D Zernike descriptors (3DZD) to represent local surface regions, facilitating quick and alignment-free shape complementarity calculations. The team adapted LZerD to use PDB files by simulating surface atoms and modified it to exclude atom-related components, raising the prefiltering threshold to expedite calculations. The resulting score matrix and PLY files accurately represented the docking poses for submission. The score matrix was prepared as a CSV file with a header row followed by rows containing the query ID and the corresponding scores for each target. Higher scores indicated better docking structures. The docking pose PLY files were generated by applying rigid rotations and translations to the target mesh using Open3D, with file names indicating the query and target IDs. This approach showed success in the CAPRI challenge [6].

- **Tran et *al.***: The team simplified the 3D models of proteins using Blender, making the models ten times less complex while keeping important details. They then rotated each simplified model in different directions to get various views. To do this evenly, they used a method called the Fibonacci Sphere. For each view, they projected the 3D model onto a flat 2D surface, focusing on capturing the highest points. This process turned each protein model into multiple simpler 2D images from different angles. They evaluated these simplified models to produce a score matrix that shows how likely each pair of protein models is to interact. This evaluation involved analyzing how well the shapes of the proteins matched up, ensuring their surfaces would fit together. They identified the top 10 most likely interactions based on these scores. Using the information from the projections, they combined the protein models into complexes to see how they fit together. To save time, they reduced the number of directions and rotations in some steps.

2.2 Dataset and Ground Truth

All structural data used to produce the dataset and ground truth utilized in this study were sourced from the Protein Data Bank [2], the Protein-Protein Docking Benchmark 5.5 [21], and the PDBFlex database [13]. The Protein-Protein Docking Benchmark 5.5 provides detailed information on protein structures that describe multichain complexes as well as their individual, non-bonded counterparts. The PDBFlex database organizes PDB structures of individual protein chains, emphasizing the structural variations and intrinsic flexibility inherent to each protein. PDBFlex works by clustering related structures from the PDB, allowing researchers to analyze and compare the conformational changes and flexibility of specific proteins.

For a given complex resolved in a single structure, the Protein-Protein Docking Benchmark 5.5 offers non-bonded counterparts for each individual chain of the complex. To streamline the docking problem within the constraints of our study timeline, we focused exclusively on complexes consisting of two protein chains. Thus, for each selected complex, we obtained three PDB structures: one representing the two-chain complex, and one for each chain in its unbound state. The PDB structure of the complex served as the ground truth,while the structures of the unbound chains were used to query PDBFlex and thereby generate different conformations of the query and target for our dataset. In the context of docking, the query and target are two separate partners of a complex.

For the queries and targets, we queried PDBFlex to determine if the chain was part of a cluster. If the chain was part of a cluster, we retrieved all the members of that cluster. If the chain was not in a cluster, the complex was removed from the dataset. We also eliminated complexes where the target cluster had fewer than 10 structures, resulting in ground truth of 52 complexes.

The main steps of this process:

1. **Identification and clustering** : We first identified if each chain was part of a cluster according to PDBFlex. This allowed us to group similar chains and understand the possible structural variations.

2. **Filtering complexes** : Complexes whose target clusters had fewer than 10 structures were removed to ensure statistical robustness in our analyses.

3. **Random selection** : To avoid selection bias and maintain diversity in our dataset,

we randomly selected up to 10 members from the cluster for both queries and targets, excluding the chain from the complex's PDB structure

4. **Balancing the dataset** : By keeping a consistent number of 10 targets for each query, we ensured that each complex was fairly represented, facilitating more reliable comparisons in docking results.

For each query, we randomly selected up to 10 members from the cluster, excluding the chain from the complex PDB structure to prevent bias. For the targets, we also randomly selected 10 members, again excluding the chain from the complex PDB structure. This approach resulted in a balanced dataset with a consistent number of 10 targets for each query. In total, the dataset that we sent to participants contained 387 queries and 520 targets and the ground-truth that we used as reference for us contained 52 complexes. The structures were then retrieved and deprotonated, and their SES were calculated using EDTSurf [22].

2.2.1 Solvent Excluded Surface (SES)

By excluding solvent molecules from the protein surface, the Solvent-Excluded Surface (SES) provides a more accurate representation of potential binding sites for other proteins
.

EDTSurf takes as input the three-dimensional structure of a molecule and produces triangular meshes that represent the SES. These meshes are typically stored in PLY files, which are a standard format for storing 3D data. A PLY (Polygon File Format or Stanford Triangle Format) file contains a description of the geometry of a 3D object. It includes a list of vertices, which are the points in space that define the shape of the object, and faces, which describe how these points are connected to form the surface. In the context of molecular structures, the vertices correspond to positions of atoms, and the faces represent the connections between these positions, effectively modeling the surface of the molecule.

The ability to accurately model the SES of proteins is crucial for understanding their interactions and designing new proteins with tailored properties. Figure 4 illustrates the application of EDTSurf in generating these surface representations.

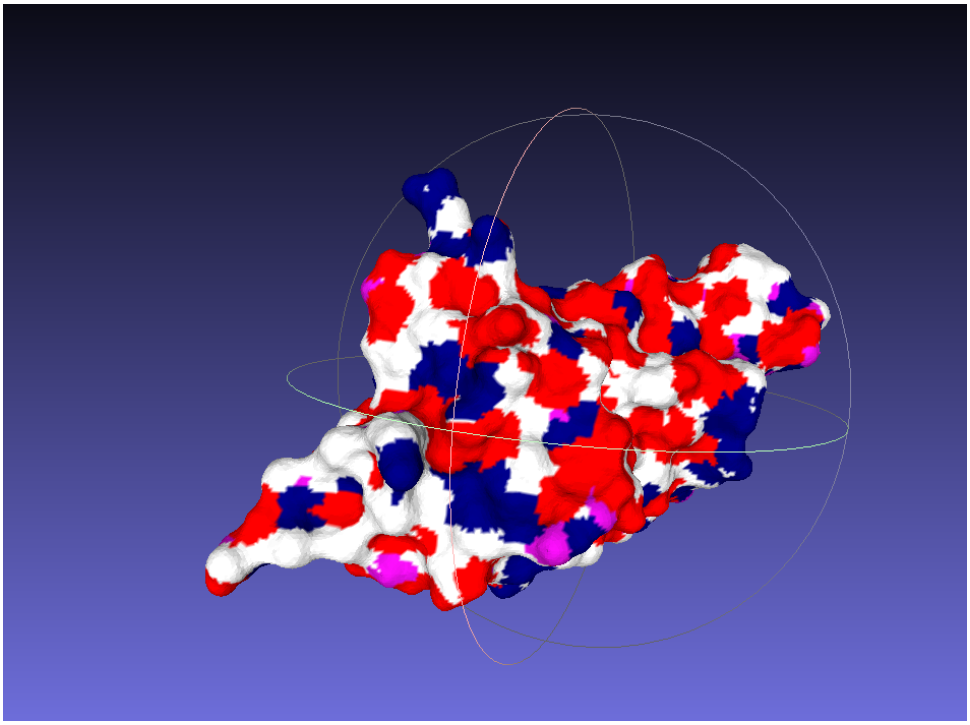

Figure 4: Example of a file PLY visualized in Meshlab [7]

.

2.3 Workflow analysis

2.3.1 Workflow

To achieve these tasks, our methods involved several key steps:

1. **Selection of True Positives**:

   - True positives are defined as the correctly predicted protein-protein complexes where the predicted interacting partners accurately match the known interacting partners of the complexes.
   - We began by identifying true positives among the top 10 query-target pairs predicted by the participants. This selection process involved evaluating the predicted complexes for their likelihood of being correct interactions, ensuring that the predictions closely matched the experimentally verified or benchmarked interactions.

2. **Alignment Using RANSAC**: To align the point clouds (PLY files) submitted by participants with the ground truth, we utilized the RANSAC (Random Sample

Consensus) algorithm [11]. RANSAC, an iterative method for estimating model parameters in the presence of outliers, was employed to find the optimal superposition and obtain the transformation matrix, specifically the rotation and translation matrix. This ensured accurate alignment of the predicted complexes with the ground truth shapes.

3. **Conversion to PDB Files**: RANSAC was used to compute the transformation matrix needed to align the query shapes from the dataset (and derived from the ground truth PDB files) to the query shapes submitted by the participants. Once this matrix was obtained, we applied it to the dataset PDB coordinates. This ensured that the same rotation and translation applied to the participants' PLY files were also applied to the corresponding PDB files. Consequently, we obtained the complexes predicted by the participants in the PDB format from the participants' submitted PLY files.

4. **Comparison Using OpenStructure**: The final step involved comparing the PDB files generated by the participants to the ground truth complexes using OpenStructure. It computes all relevant scores in a single run, providing a comprehensive evaluation of the predicted complexes (cf. § 2.4.1).

2.3.2 Tools and Techniques

- **RANSAC**: RANSAC (Random Sample Consensus) is a robust iterative method used to estimate the parameters of a transformation model from a set of observed data that may contain outliers. In the context of 3D alignment, RANSAC was employed to find the optimal superposition of point clouds. This method is robust against outliers and ensures that the best alignment is achieved by iteratively fitting the transformation model to subsets of the data and selecting the model that best represents the entire dataset. The process begins with initial sampling, where a small subset of the data points is randomly selected. This is followed by model fitting, where a transformation model is fitted to this subset. Next, consensus set identification takes place, where the model is evaluated against the entire dataset to identify points that fit within a predefined tolerance, known as inliers. The model is then refined using all identified inliers, and this process is repeated to find the

model that maximizes the number of inliers.

- **OpenStructure**: OpenStructure was utilized for the final comparison of PDB files [3]. Its capability to compute all necessary scores (lDDT, ICS, IPS, DockQ, TM-score) in a single run greatly facilitates reproducibility. One of the key strengths of OpenStructure is its ability to compute different scores, including : Root Mean Square Deviation (RMSD) which measures the average distance between atoms of superimposed proteins' structures Interface RMSD (iRMSD), which specifically measures the RMSD at the interface region, where two proteins interact. Fraction of Common Contacts (FCC), which assesses the similarity in the contact patterns of the interface residues, indicating the spatial proximity between atoms involved in potential interactions. This analysis typically considers contacts within a specified distance threshold, such as a cutoff distance of 5 angstroms. OpenStructure automatically cleans up loaded structures by assigning elements based on the PDB Chemical Component Dictionary, removing hydrogens and unknown atoms. The cleaned data is output in JSON format, including chain names, chemically equivalent groups of polypeptides, mappings of model chains to these groups, and pairwise sequence alignments in FASTA format.

2.4 Evaluation

2.4.1 CAPRI scoring

CAPRI (Critical Assessment of PRedicted Interactions) is a community-wide experiment aimed at predicting the structures of macromolecular complexes, with a particular focus on protein-protein interactions. It provides a rigorous framework for evaluating the accuracy of predicted models by comparing them against experimentally determined reference structures. In our study, we utilized CAPRI's established metrics to assess the quality of our docking predictions. These metrics allowed us to objectively measure the performance of the participants' models in terms of how well they predicted the interactions and conformations of protein complexes. By adhering to CAPRI's standards, we ensured that our evaluation was aligned with widely accepted benchmarks in the field of structural bioinformatics. Scores used during our challenge are:

- ICS (Interface Contact Similarity) [15]

- IPS (Interface Patch Similarity) [15]
- LDDT (Local Distance Difference Test) [19]
- TM-score(Template Modeling score) [23]
- DockQ [1]

Interface Contact Similarity (ICS) : Interface Contact Similarity (ICS) is an F1-score that quantifies the similarity between sets of predicted and native contact points at the protein-protein interface. ICS identifies specific atomic contacts at the interface, typically defined by a distance threshold.

$$\textbf{ICS}(\textbf{M},\textbf{T}) = \text{F1(P,R)} = 2 \cdot \frac{P(M,T) \cdot R(M,T)}{P(M,T) + R(M,T)}$$

Where:

- **M** represents the model protein complex.
- **T** represents the target (reference) protein complex.
- **P** Precision) is the fraction of correct inter-chain contacts among all predicted inter-chain contacts.
- **R** (Recall) is the fraction of correctly reproduced native inter-chain contacts.

Two residues are considered to be in contact if at least two of their non-hydrogen atoms (one from each residue) are within 5Å of each other. Precision (P) measures how many of the predicted contacts are correct and recall (R) measures how many of the native contacts are correctly predicted.

Interface Patch Similarity (IPS)

Interface Patch Similarity (IPS) is calculated as a Jaccard coefficient (Jc) over the interface amino acids (I) predicted by the model (I_M) compared to the target (I_T). The formula for IPS is:

$$\text{IPS}(M,T) = J_C(I_M, I_T) = \frac{|I_M \cap I_T|}{|I_M \cup I_T|} \tag{1}$$

Where:

- **M**: represents the model protein complex.

- **T**: represents the target (reference) protein complex.
- $\boldsymbol{I_M}$: denotes the set of interface amino acids predicted by the model.
- $\boldsymbol{I_T}$: denotes the set of interface amino acids in the target.
- **Intersection** ($\cap$): The common amino acids between the predicted and target interfaces.
- **Union** ($\cup$): The total amino acids present in either the predicted or target interfaces.

The Jc measures the similarity between the sets of predicted and native interface residues. It ranges from 0 to 1, where 0 indicates no overlap and 1 indicates perfect overlap.

Local Distance Difference Test (lDDT) : lDDT quantifies the degree of structural similarity between the predicted protein structure and a reference or experimental model by analyzing the preservation of inter-atomic distances between corresponding atoms in the model and the reference structures [11]. To comprehensively evaluate structural fidelity, LDDT employs multiple thresholds, typically including distances of less than 0.5 Å, 1 Å, 2 Å, and 4 Å. The final LDDT score is calculated as the average of four fractions, with each fraction representing the proportion of preserved distances at the respective threshold.

Template Modeling score: TM-score evaluates the overall alignment of the predicted structure with the reference structure, considering both backbone and side-chain atoms. It is expressed as:

$$TM\text{-}score = \max \left[\frac{1}{L_N} \sum_{i=1}^{L_T} \frac{1}{\left(1 + \left(\frac{d_i}{d_0}\right)^2\right)} \right] \tag{2}$$

where:

- L_N is the length of the native structure, representing the total number of residues in the reference or experimental protein structure.
- L_T is the length of the aligned residues to the template structure, indicating the

number of residues that can be successfully aligned between the predicted and reference structures.

- d_i is the distance between the ith pair of aligned residues, representing the spatial separation between corresponding residues in the two structures after optimal spatial superposition.

- d_0 is a scale factor used to normalize the match difference, ensuring that the ™-score remains within the range of (0, 1]

DockQ: DockQ estimates the deviation of predicted protein-protein complex structures from experimental data using a Z-score. DockQ evaluates how closely the predicted protein-protein complex structures match experimental data, considering factors such as binding affinity and interface geometry. It calculates a Z-score to statistically compare the predicted structures against experimental benchmarks.

On the other hand, ICS, IPS, and DockQ focus on assessing the retrieval of the interaction interface between both partners of the complex. These metrics evaluate the accuracy of predicted binding interfaces and the overall quality of the protein-protein interaction predictions. On the other hand, lDDT and TM-score reflect the overall similarity between the ground truth complex and those submitted by the participants.

The values for all scores are typically bounded between 0 and 1.

2.4.2 Other metrics

In evaluating the retrieval performances of participants' algorithms in the SHREC challenges, a comprehensive array of metrics was employed to judge the efficacy of their methods. These metrics, including standard Nearest-Neighbor (NN), First-Tier(FT), and Second-Tier(ST) retrieval evaluations, allow evaluation of the algorithmic performance. Specifically, these metrics quantify the fraction of true positives, reflecting the accuracy of each query's results across different tiers of retrieval. For instance, the NN metric measures the fraction of true positives within the top-ranked result. Similarly, the FT metric extends this evaluation to the top-N results, where N represents the number of true positives in the ground truth for the considered query. Furthermore, the ST metric offers additional evaluation by assessing the performance across 2N first result).

In our analysis, we alsocalculated the vectors between the center of masses of the query and target proteins in both the predicted poses and reference structures, following a query-based superposition of the predicted poses. By measuring the angle between these vectors, we obtained insights into the extent of alignment or divergence in the target's position relative to the query in both the predicted poses and reference structures.

Results

3.1 Retrieving results

3.1.1 True positive rates

We evaluated the true positive rates at both the query and protein levels to assess the effectiveness of the participants' algorithms in identifying correct protein-protein interactions. True Positive (TP) refers to the correct prediction of protein-protein interactions by the participants' algorithms, calculated as the number of true positive interactions among all the interactions predicted by the participants for a given set of protein complexes. True Positive Rate at the Protein Level (TPcomplex) metric reflects the percentage of protein complexes, out of the total of 52 complexes in the ground truth dataset, for which at least one query resulted in a true positive interaction. These results are presented in Table 2.

Metric	**FS-proj**	**LZerD**
TP(TP - %)	49(1.3%)	99(2.6%)
TP_{complex} (TP_{complex} − %)	22(42.3%)	12(23.1%)

Table 2: True Positive Rates

The FS-proj method, compared to the LZerD method, identified fewer true positive matches per complex.

3.1.2 Performance of the Evaluated Algorithms

In addition to true positive rates, we used standard retrieval metrics to further evaluate the performance of the participants' algorithms: Nearest-Neighbor (NN), First-Tier (FT), and Second-Tier (ST). These results are presented in Table 3:

NN specifically checks if the very first result returned by the algorithm matches the correct interaction. For both FS-proj and LZerD, the NN score was 0, indicating that the

Metric	FS-proj	LZerD
Nearest-Neighbor	0.0	0.0
First-Tier	0.014	0.023
Second-Tier	0.035	0.046

Table 3: Nearest-Neighbor (NN), First-Tier (FT), and Second-Tier (ST) Metrics

closest results did not include any true positives for either algorithm. FT metric measures the proportion of true positives found within the top results up to the size of the query set. It provides an indication of how well the algorithm performs when considering a limited number of top-ranked results. FS-proj had an FT score of 0.014, while LZerD had an FT score of 0.023, suggesting that LZerD was slightly more effective at finding true positives within the top tier of results. ST metric extends the evaluation to include twice the size of the query set, providing a broader measure of retrieval performance beyond the immediate top results. It shows how well the algorithm performs when given a larger set of potential true positives. FS-proj had an ST score of 0.035, and LZerD had an ST score of 0.046, indicating that both algorithms performed better when allowed a wider range of results, with LZerD again showing a slight advantage.

3.1.3 CAPRI scoring

Figure 5 displays the distribution of key scoring metrics, including ICS, IPS, DockQ, TM-score, and lDDT, for each true positive identified by the participating teams in our study. Interestingly, similar trends were observed in the results obtained by both teams. The results shown in Table 1 and Figure 5 highlight some challenges both methods faced in accurately identifying the correct protein interactions. For both methods, the ICS, IPS, and DockQ scores are below 0.1, indicating limited agreement in the predicted interaction interfaces when compared to the ground truth. These low scores suggest that the predicted locations of protein interactions do not closely match the actual interaction sites observed in the reference structures. In contrast, lDDT and TM-score exhibit significantly higher values, ranging between 0.5 and 0.9, suggesting a closer structural alignment between the predicted and reference complexes.

The differences in performance between these metrics can be attributed to their distinct evaluation criteria. ICS, IPS, and DockQ focus on assessing the retrieval of the interaction interface between the two partners of the complex, while lDDT and TM-score

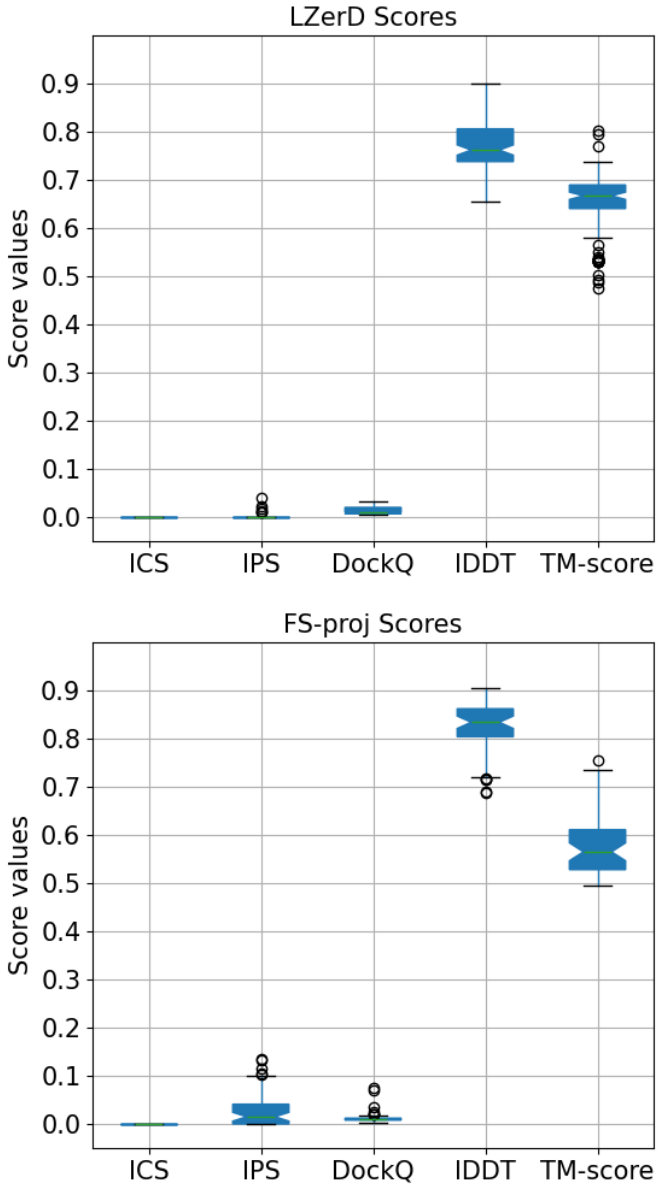

Figure 5: Evaluation of result quality compared to ground truth. Five parameters were assessed for each participant: LZerD (upper) and FS-proj (lower), including ICS, IPS, DockQ, lDDT, and TM-score. Scores range from 0 to 1. Overall, ICS, IPS, and DockQ exhibit low scores (between 0 and 0.2), while TM-score and lDDT exhibit high scores (between 0.5 and 0.9). .

measure the overall structural similarity between the ground truth complex and the models submitted by the participants. As a result, there are instances where low ICS, IPS,

and DockQ scores coexist with high lDDT and TM-score values. This indicates partial alignment in certain regions of the complex, typically seen in the query structures, despite incomplete retrieval of the interaction interface.

3.1.4 Vector scoring

Although IPS and ICS are used to compare specific interaction interfaces between proteins, we also a third vector-based criterion was employed to evaluate the accuracy of the predicted interaction zones. Unlike ICS and IPS, which focus on the comparison of contacts at the interface, the vector-based criterion assesses the spatial relationship between the query and target proteins.

This method involves comparing the vector between the centers of mass of the query and target in each predicted model to the corresponding vector in the reference structure. This comparison provides a measure of how accurately the predicted position of the target matches its actual position relative to the query.

The results, presented in Figure 6, show that approximately 20% of the predicted complexes from both participants correctly positioned the target within a 60° threshold around the reference target's location. Additionally, considering angles below 30°, FS-proj outperformed other methods, indicating a slightly better performance in predicting complexes with interfaces closer to the true interface. This observation suggests that the methods employed by the participants, while not always precisely predicting the reference interface, are capable of approximating the interaction zone near the actual interface.

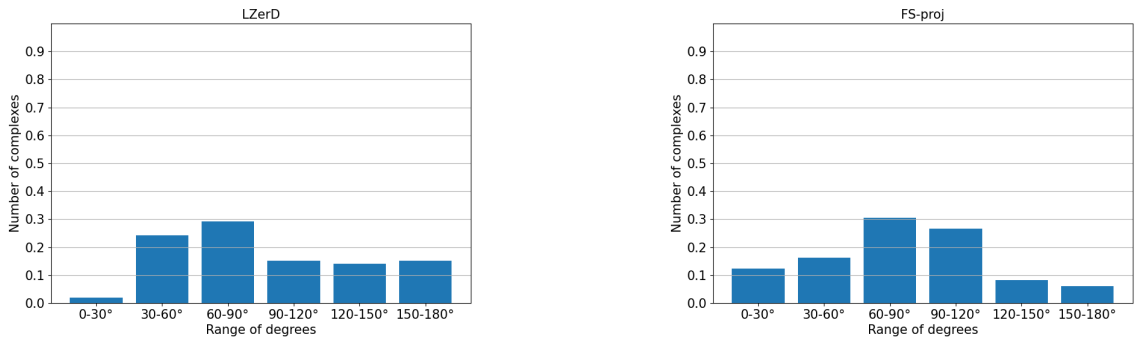

Figure 6: Proportion of complexes relative to the angle between center-of-mass vectors.

Discussion

4.1 True positive rates

As shown in Table 2, the LZerD method achieved a higher overall true positive rate, with 99 true positives (2.6%), compared to FS-proj's 49 true positives (1.3%). This indicates that LZerD correctly identified more interacting proteins overall. However, when evaluating the true positive rate per complex (TPcomplex), FS-proj demonstrated broader coverage, identifying true positives in 22 complexes (42.3%), compared to LZerD's 12 complexes (23.1%). To further understand the effectiveness of each method, we calculated the average number of true positives per complex (TP/TPcomplex). For LZerD, this metric is 8.25, derived from dividing the 99 true positives by the 12 complexes where at least one true positive was identified. This high average suggests that LZerD was very effective at finding multiple correct interactions within the complexes it predicted. In contrast, FS-proj has an average of 2.23 true positives per complex, calculated by dividing 49 true positives by 22 complexes. This lower average indicates that while FS-proj identified fewer true positives within each complex, it successfully found at least one true positive in a larger number of different complexes. This additional metric (TP/TPcomplex) helps to highlight the strengths and weaknesses of each method. LZerD's higher average indicates depth in its predictions, finding multiple interactions within fewer complexes. FS-proj's broader coverage across more complexes, despite a lower average per complex, suggests it is better at identifying at least one correct interaction in a wider range of complexes.

4.2 CAPRI scoring

Additionally, the quality of the complexes identified by FS-proj was comparable, with ICS (values = 0) and DockQ (values between 0 and 0.1) scores remaining equally low for both methods.There are slight differences observed for the IPS metric, with FS-proj exhibiting slightly higher values, albeit still below 0.1, as illustrated in Figure 5. IPS measures how well the predicted interaction interfaces align with the actual interfaces. These metrics assess the accuracy of the predicted interaction interfaces compared to the actual interfaces in the reference data. Low scores in these metrics suggest that even when the correct query-target pair was identified, the detailed relative position and/or

orientation of the target protein in relation to the query protein was often inaccurate. This discrepancy is further highlighted by the distribution of angles between the vectors connecting the center of mass of the query protein to the center of mass of the target protein, as shown in Figure 6.

4.3 Vector scoring

Moreover, FS-proj predicted more poses (orientations of the proteins) where the angle between the vector from the center of mass of the target protein to the center of mass of the query protein in the predicted pose and the same vector in the actual complex was less than 30°. This means that these complexes are predicted to contain a better interface. The closer this angle is to 0°, the more aligned the vectors are, indicating a more accurate prediction. Observing more complexes with vector angles less than 30° signifies that FS-proj predicts more high-quality complexes since an angle below 30° means the predicted interface is closer to the true interface. The FS-proj method exhibited slightly superior complex quality for the true positives, as indicated by marginally higher IPS values (see Figure 5).

4.4 LZerD Adaptation for Shape-based Docking in SHREC2024

The LZerD method was originally developed to work with detailed protein structures and has been proven successful in predicting protein complexes when the structures of the individual protein partners are known. This method typically operates using Protein Data Bank (PDB) files, which contain detailed 3D structural information about the proteins. In our SHREC2024 track, which focuses on shape complementarity, the challenge was to adapt LZerD to work with surface meshes instead of detailed protein structures. To adapt this structure-based algorithm to the shape-based problem posed by SHREC2024, the surface meshes (which represent the outer shape of the proteins) had to be converted back into synthetic PDB structures. This conversion process involves generating a simplified 3D model of the protein that approximates its surface shape. These synthetic structures were then subdivided into smaller local surface areas to facilitate docking, which is the process of predicting how two proteins interact. However, this conversion from surface mesh to synthetic PDB structure likely introduced inaccuracies. The synthetic structures may not have been precise enough to allow the LZerD docking algorithm to produce high-

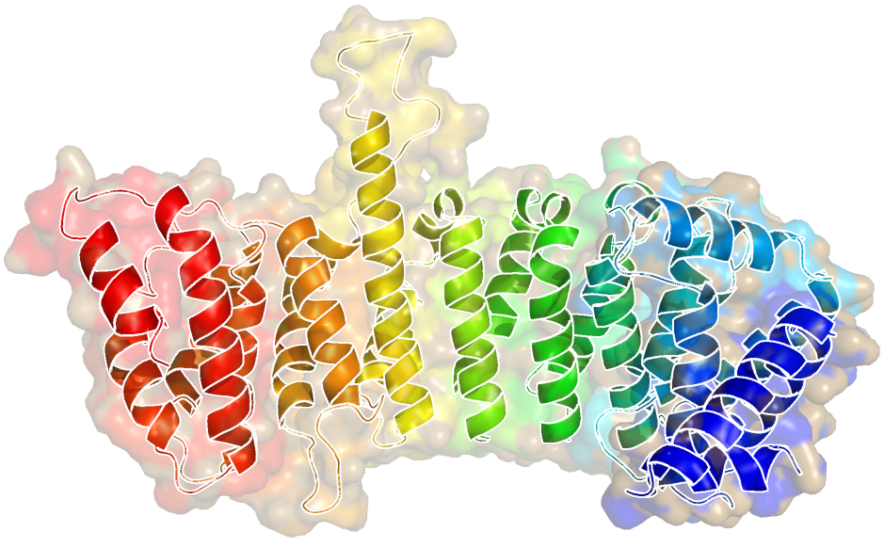

Figure 7: Structure of Importin beta-1 subunit with the rainbow coloring representing the various secondary structures, predominantly composed of approximately twenty alpha helices.

quality predictions. This is supported by the performance gap observed between this SHREC2024 track and previous CAPRI experiments. In CAPRI, where LZerD operates directly on detailed protein structures, the method has consistently produced high-quality predictions [5]. The additional step of converting surface meshes to PDB structures in SHREC2024 may have compromised the structural fidelity necessary for optimal docking performance. A critical component of the LZerD method is its scoring function, which determines how well two proteins (referred to as the query and the target) fit together. This scoring function evaluates several factors to rank the possible ways in which the proteins might interact. One important factor is the area of the interface where the surfaces of the two proteins come into contact. Generally, a larger interface area suggests a more stable and likely interaction because it indicates more extensive contact between the two protein surfaces.

To illustrate, let's consider the complex formed by the Importin beta-1 subunit (Figure 7) and the GTP-binding nuclear protein RAN (Figure 8). The Importin beta-1 subunit exhibits an irregular crescent shape, resulting in a concave site that offers a substantial surface area for potential interaction (refer to Figure 9). This extended contact region is

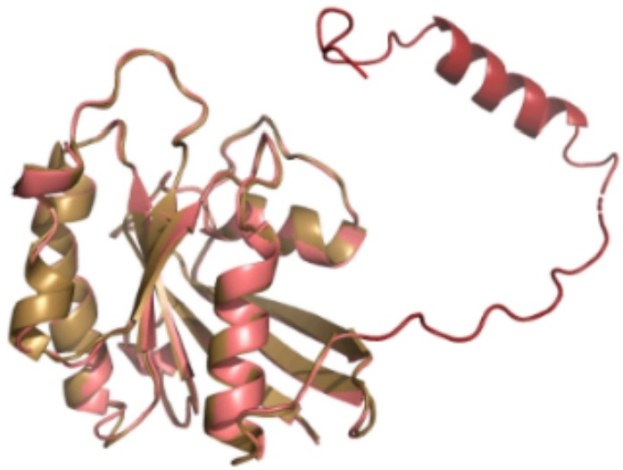

Figure 8: Comparison of RAN Protein Conformations: Ground Truth (yellow) vs Dataset (pink)

particularly advantageous for docking predictions, suggesting a more stable interaction with the target protein. The solvent accessible surface area (SASA) of the Importin beta-1 subunit is measured at 22,735 Å^2. Comparatively, the average SASA across all proteins (sub-units) within the ground truth dataset stands at approximately 10,664 Å^2, with a standard deviation of 6,471 Å^2. These values provide concrete evidence, illustrated in the figure 7, of the Importin beta-1 subunit's significantly larger contact area. Despite some challenges with the initial step in the LZerD workflow, where the surface meshes are converted to synthetic protein structures, the overall complementary shapes of the Importin beta-1 subunit and the GTP-binding nuclear protein RAN were still effectively identified. This particular protein complex was accurately predicted 39 times, making it the most frequently identified true positive in the dataset. The strong shape complementarity between these two proteins, combined with a scoring function that favors larger interface areas, likely explains why this specific complex was over-represented among the successful predictions.

4.5 FS-proj methodology in protein docking

The FS-proj method identifies matching surfaces between query and target proteins using a multi-view approach. Initially, the detailed 3D shapes of the proteins, known as

meshes, are simplified by removing 90% of the meshes' points This simplification reduces computational load and speeds up the subsequent steps. Following this, the method generates multiple perspectives of each protein's surface by rotating and displacing the simplified 3D shapes in various directions, creating different viewpoints of the surface geometry. Each of these 3D views is then projected onto a 2D plane, converting the three-dimensional information into a 100 × 100 grid or table. These tables serve as maps that encode the surface geometry of the protein from specific angles. After the 2D tables are created, they are aligned to compare the surfaces of the query and target proteins. A distance metric measures how well these tables fit together, indicating the degree of match between the protein surfaces. The FS-proj method depends on the accurate preservation of crucial surface details throughout the simplification, multi-view generation, and 2D projection steps to effectively identify complementary interfaces between proteins.

4.6 Common analysis of both teams

Both methods showed low scores for accurately predicting the specific areas where the proteins interact (measured by ICS, IPS, and DockQ scores). These scores indicate how well the predicted interaction sites match the actual sites. On the other hand, the methods performed better when assessing the overall shape alignment between the predicted and reference protein structures (measured by lDDT and TM-score), which were significantly higher. Another challenge noted was the accuracy in predicting the relative positions and orientations of the target proteins in relation to the query proteins. The metrics used to align the 3D representations of protein surfaces could be improved.

Finally, creating a dataset for our study had its challenges, particularly when comparing protein shapes between our dataset and the original data source. This complexity explains the failure of our approach, as the dataset does not include the ground truth conformations but rather "unbound" conformations that are not necessarily optimal. Consider the RAN protein (Figure 8) as an example. Imagine this protein can take on different shapes, even though it's the same protein. One end of the RAN protein is highly flexible and might not be well-documented in most of our data sources, like the Protein Data Bank (PDB). That's because the tools we used to study proteins struggle to capture these flexible parts accurately. However, in certain instances, when the RAN protein interacts with specific partners, this flexible part might adopt a more stable shape that

can be recorded in the data sources.

Their methodologies need to be better adapted. However, this lack of adaptation during the challenge can be explained by the time constraint, which is why participants did not have much time to better adapt their methodology.

Conclusion and Perspective

During the work on this project, we have outlined a challenging endeavor focused on the retrieval of non-rigid shape complementarity in protein-protein interactions. This task had some difficulties such as stringent time constraints, intricate protein structures and presence of significant rearrangements within the dataset. Despite this, both methods showed low scores for accurately predicting the specific areas where the proteins interact (measured by ICS, IPS, and DockQ scores). These scores indicate how well the predicted interaction sites match the actual sites. On the other hand, the methods performed better when assessing the overall shape alignment between the predicted and reference protein structures (measured by lDDT and TM-score), which were significantly higher. Another challenge noted was the accuracy in predicting the relative positions and orientations of the target proteins in relation to the query proteins. The metrics used to align the 3D representations of protein surfaces could be improved.

Finally, creating a dataset for our study had its challenges, particularly when comparing protein shapes between our dataset and the original data source as we demonstrated with the example of the RAN protein, as discussed.

Our examination of two participant approaches, LZerD by Kihara et al. and FS-proj by Tran et al., has shed light on both the potential and the limitations of current methodologies. Despite the innovative strategies employed by participants, the mixed results were evident. However, amidst these challenges lies a significant opportunity for future research and development. Our collaborative effort serves as a foundational step towards the advancement of protein shape docking methods, offering a curated dataset that is publicly available and ripe for further expansion and optimization. The outcomes of this collaborative effort will be meticulously compiled into a comprehensive report, ensuring the thorough documentation and dissemination of our findings within the scientific community. Additionally, the results will be presented at an upcoming symposium,

providing a forum for in-depth discussion, feedback, and collaboration among researchers.

In SHREC2025, the focus extends beyond protein-protein docking to include surface-based analysis for remote protein classification, aiming to identify distantly homologous proteins with similar shapes but different sequences. This builds on the work from previous SHREC challenges, such as SHREC2021 and SHREC2019, which concentrated on classification tasks. These challenges emphasized the importance of shape similarity in identifying proteins with related functions despite significant differences in their primary sequences. Currently, our team is collaborating on the Gradient project with a team from LIRIS and CGI, working on classification algorithms using machine learning techniques to further enhance the accuracy and efficiency of protein classification. You can find more information about the Gradient project on their website: https://gradient.projet.liris.cnrs.fr/.

References

[1] S Basu and B Wallner. Dockq: A quality measure for protein-protein docking models. *PLoS One*, 11(8):e0161879, 2016.

[2] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, and P.E. Bourne. The protein data bank. *Nucleic Acids Research*, 28:235–242, 2000.

[3] M. Biasini, T. Schmidt, S. Bienert, V. Mariani, G. Studer, J. Haas, N. Johner, A. D. Schenk, A. Philippsen, and T. Schwede. Openstructure: an integrated software framework for computational structural biology. *Acta Crystallographica Section D: Biological Crystallography*, 2013.

[4] A. v. Bondi. van der waals volumes and radii. *The Journal of Physical Chemistry*, 68(3):441–451, 1964.

[5] C. Christoffer, G. Terashi, W. Shin, T. Aderinwale, S. R. Venkata Subramaniya, and L. Peterson. Performance and enhancement of the lzerd protein assembly pipeline in capri. *Proteins: Structure, Function, and Bioinformatics*, 88(8):948–961, 2019.

[6] C Christoffer, G Terashi, WH Shin, T Aderinwale, SR Maddhuri Venkata Subramaniya, L Peterson, J Verburgt, and D Kihara. Performance and enhancement of the lzerd protein assembly pipeline in capri 38-46. *Proteins*, 88(8):948–961, 2020.

[7] P. Cignoni, M. Callieri, M. Corsini, M. Dellepiane, F. Ganovelli, and G. Ranzuglia. Meshlab: an open-source mesh processing tool. In *Sixth Eurographics Italian Chapter Conference*, pages 129–136, 2008.

[8] M. L. Connolly. Shape complementarity at the hemoglobin 11 subunit interface. *Biopolymers*, 25(7):1229–1247, 1986.

[9] Javier De Las Rivas and Celia Fontanillo. Protein-protein interactions essentials: key concepts to building and analyzing interactome networks. *PLoS Comput Biol*, 6(6):e1000807, 2010.

[10] H. Dutagaci, A. Godil, A. Axenopoulos, P. Daras, T. Furuya, and R. Ohbuchi. Shrec'09 track: querying with partial models. In *Proceedings of the 2nd Eurographics*

Conference on 3D Object Retrieval (3DOR '09), pages 69–76, Goslar, DEU, 2009. Eurographics Association.

[11] Martin A. Fischler and Robert C. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM*, 24:381–395, 1981.

[12] Pablo Gainza, Freyr Sverrisson, Filippo Monti, et al. Deciphering interaction fingerprints from protein molecular surfaces using geometric deep learning. *Nat Methods*, 17:184–192, 2020.

[13] Thomas Hrabe, Zhanwen Li, Mayya Sedova, Piotr Rotkiewicz, Lukasz Jaroszewski, and Adam Godzik. Pdbflex: exploring flexibility in protein structures. *Nucleic Acids Research*, 44(D1):D423–D428, 2016.

[14] 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*, 52(1):2–9, Jul 2003.

[15] Aleixandre Lafita, Spencer Bliven, Andriy Kryshtafovych, Martino Bertoni, Bohdan Monastyrskyy, José M Duarte, et al. Assessment of protein assembly prediction in casp12. *Proteins: Structure, Function, and Bioinformatics*, 86(S1):247–256, 2018.

[16] F. Langenfeld, T. Aderinwale, C. Christoffer, W. H. Shin, G. Terashi, X. Wang, D. Kihara, H. Benhabiles, K. Hammoudi, A. Cabani, F. Windal, M. Melkemi, E. Otu, R. Zwiggelaar, D. Hunter, Y. Liu, L. Sirugue, H. H. Nguyen, T. H. Nguyen, V. T. Nguyen-Truong, D. Le, H. D. Nguyen, M. T. Tran, and M. Montès. Surface-based protein domains retrieval methods from a shrec2021 challenge. *Journal of Molecular Graphics and Modelling*, 111:108103, Mar 2022. Epub 2021 Dec 21.

[17] B. Lee and F. M. Richards. The interpretation of protein structures: estimation of static accessibility. *Journal of Molecular Biology*, 55(3):379–IN4, 1971.

[18] Mohamed Machat, Florent Langenfeld, Daniela Craciun, Léa Sirugue, Taoufik Labib, Nathalie Lagarde, Maxime Maria, and Matthieu Montes. Comparative evaluation of shape retrieval methods on macromolecular surfaces: an application of computer

vision methods in structural bioinformatics. *Bioinformatics*, 37(23):4375–4382, December 2021.

[19] Valerio Mariani, Marco Biasini, Alessandro Barbato, and Torsten Schwede. lddt: A local superposition-free score for comparing protein structures and models using distance difference tests. *Bioinformatics*, 2013.

[20] Chaoyu Quan, Benjamin Stamm, and Yvon Maday. A domain decomposition method for the polarizable continuum model based on the solvent excluded surface. *Mathematical Models and Methods in Applied Sciences*, 28(07):1233–1266, 2018. _eprint: https://doi.org/10.1142/S0218202518500331.

[21] T. Vreven, I. H. Moal, A. Vangone, B. G. Pierce, P. L. Kastritis, M. Torchala, R. Chaleil, B. Jiménez-García, P. A. Bates, J. Fernandez-Recio, A. M. Bonvin, and Z. Weng. Updates to the integrated protein-protein interaction benchmarks: Docking benchmark version 5 and affinity benchmark version 2. *Journal of Molecular Biology*, 427(19):3031–3041, Sep 2015.

[22] Xueqing Xu, Moses M. Kasembeli, Xiaoyu Jiang, Benjamin J. Tweardy, and David J. Tweardy. Chemical probes that competitively and selectively inhibit stat3 activation. *PLoS ONE*, 4(3):e4783, 2009.

[23] Yang Zhang and Jeffrey Skolnick. Scoring function for automated assessment of protein structure template quality. *Proteins: Structure, Function, and Bioinformatics*, 57(4):702–710, 2004.