Identifying Candidates for Protein-Protein Interaction: A Focus on NKp46's Ligands

Alessia Borghini[1,*], Federico Di Valerio[1,*], Alessio Ragno[1] and Roberto Capobianco[2]

[1] *Sapienza University, Rome*

[2] *Sony AI*

Abstract

Recent advances in protein-protein interaction (PPI) research have harnessed the power of artificial intelligence (AI) to enhance our understanding of protein behaviour. These approaches have become indispensable tools in the field of biology and medicine, enabling scientists to uncover hidden connections and predict novel interactions. The experimental processes to analyze and validate the interactions between proteins are usually expensive and time-consuming and with this work, we can reduce these costs by strategically filtering and computationally validating the possible proteins which might take part in the interactions at hand. Aiming at helping in broadening the repertoire of known interacting proteins, we present a method for the systematic screening of proteins that exhibit a high affinity for the interaction with a chosen protein. Specifically, building upon already known protein interactions, we exploit the self-explainability of the deep learning model DSCRIPT to search and find promising protein candidates for a determined PPI. We analyze and rank the candidates using various strategies, and then employ AlphaFold2 to validate the resulting interactions. Consequently, we compare our AI-driven methodology with traditional bioinformatics approaches commonly used to find potential protein candidates. Throughout the overall process, explanatory data is obtained, among which is an informative contact map that elucidates the potential interaction between a protein of the known interaction and the predicted proteins. As a case study, we apply our method to deepen our understanding of NKp46's ligands repertoire, which is yet not fully uncovered.

Keywords

Protein-protein interaction, Explainable AI, Natural killer cell

1. Introduction

Predominant computational models employ data-driven algorithms that assess pairs of proteins, evaluating their likelihood of interaction based on their primary features, thereby categorizing the pairs as either interacting or non-interacting. Unraveling protein-protein interactions (PPIs) represents a fundamental challenge in bioinformatics. Despite the extensive research efforts dedicated to identifying PPIs, a significant discrepancy persists between the number of experimentally verified interactions and the vast array of PPIs in biological systems.

The fraction of PPI networks that have been experimentally mapped is minimal, primarily due to the prohibitive costs and extensive time investments required by traditional experimental methodologies. Consequently, the deployment of high-throughput computational strategies becomes indispensable for the systematic discovery of protein interactions. When employed

EXPLIMED - First Workshop on Explainable Artificial Intelligence for the medical domain - 19-20 October 2024, Santiago de Compostela, Spain

ragno@diag.uniroma1.it (A. Ragno)

alongside experimental techniques, these computational-based techniques substantially elevate the fidelity and precision of PPI predictions. PPI models forecast potential binding between pairs of protein complexes, starting from their amino acid sequences. Among state-of-the-art methods, some studies [1, 2, 3, 4] propose using AlphaFold to fold the dimer of the pair and predict contact points, providing accurate results. However, while it is very accurate, predicting such interaction generally takes a moderate amount of time, posing challenges for identifying novel interacting proteins. Recognizing these needs, our study concentrates on the strategical identification of potential interacting protein candidates, informed by experimentally pre-established interaction data.

Avoiding the indiscriminate search for protein pairs with potential for interaction, our methodology exploits the explanation of deep neural networks to pinpoint proteins with a high propensity for interaction with another chosen protein. This original and simple approach initiates with the choice of a protein pair known to interact experimentally, utilizing one protein as a template to model interaction with its counterpart. The model protein undergoes a computational process, subsequently serving as a basis to identify other proteins exhibiting analogous interaction potential.

To reduce the overall process time, we propose to use a deep learning interpretable sequence-based, structure-aware, genome-scale protein-protein interaction model, DSCRIPT [5, 6], to seek potential candidates that could take part in the interaction with a pre-chosen protein. In this way, we optimize the starting phase by choosing only the candidates with the highest potential for interaction. DSCRIPT[1] is an attention-based PPI model that can be explained through visualization of attention scores used to estimate the interaction between two proteins (more about in section 2). Indeed, DSCRIPT generates contact maps of the predicted interaction that are coherent with the ground truth [5]. This inherent interpretability of DSCRIPT allows us to better strategize how to filter the potential protein candidates for interaction with a chosen protein. The high potential proteins found through the inner representations computed by DSCRIPT (obtained with different strategies) were then validated by using both DSCRIPT itself and the SpeedPPI[2] model [7], which is an innovative optimized protein structure prediction method based on the AlphaFold2 (AF2) [8] model. This step was crucial in confirming the biological relevance of the interactions identified by DSCRIPT.

Our case study is the recently discovered interaction between the NKp46 of natural killer (NK) cells with calreticulin (CRT) [9]. We focus on finding proteins similar to CRT which could potentially interact with NKp46. As part of the innate immune system, NK cells play a crucial role in detecting and eliminating infected, transformed, and stressed cells [10, 11, 12]. Among the natural cytotoxicity receptors (NCRs) that activate NK cells, NKp46 stands out for its ability to recognize and target a broad spectrum of tumour cells [13, 14, 9]. NKp46 is encoded by the NCR1 gene and is the most evolutionarily ancient NK cell receptor, expressed by most NK cells and some innate lymphoid cells. This remarkable ability is related to the receptor's interactions with its ligands. However, the specific identity of these ligands remains unclear [15]. Its activation is associated with destroying cancer cells, making it a potential target for cancer immunotherapy. However, the scarcity of identified ligands for NKp46 hinders its therapeutic

[1] https://github.com/samsledje/D-SCRIPT?tab=readme-ov-file
[2] https://github.com/patrickbryant1/SpeedPPI

application.

In this work, we specifically aim to identify potential proteins that might interact with NKp46 utilizing PPI models' inherent interpretability and inner representations.

Overall, we provide the following contributions:

- Development of a systematic screening method for proteins that exhibit a high affinity for interaction with a chosen protein, leveraging the self-explainability of the deep learning model DSCRIPT;
- Utilization of AlphaFold2 to validate the interactions predicted by DSCRIPT, comparing the AI-driven methodology with traditional bioinformatics approaches;
- Finding potential ligands for NKp46 that may play an important role in the human immune system.

The remainder of this work is organized as follows: Section 2 introduces a literature review of methods related to PPI; in Section 3 we describe in detail DSCRIPT and SpeedPPI, on which we base our work, and present our proposed approach; Section 4 provides the experimental procedures to validate our approach using NkP46 as a case study; finally, in Section 6 we wrap up our results, presenting the limitations and future directions of our work.

2. Related Work

Protein-Protein Interactions (PPIs) boast of several works leveraging prior knowledge from known interacting protein sequences and employing machine-learning (ML) techniques [16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27]. Some approaches focus solely on amino acid sequences, employing strategies such as counting amino acid triplets [19, 23, 26], defining signatures as sets of subsequences [20], assessing auto-correlation values of physicochemical scales [24, 28], or analyzing normalized counts of amino acid residues or pairs [25]. These sequence-based methods have demonstrated prediction accuracies ranging from 70% to 84% on human datasets and around 70% on yeast datasets. Additionally, several methods incorporate information about protein domains [29, 30], which has proven to be informative for PPI prediction [27]. However, domain-based methods are limited in applicability to proteins lacking domain assignments.

Identifying homologous proteins is a common strategy for inferring the functions of newly discovered proteins, as homologs typically share similar functions and three-dimensional structures. This deductive approach has also been employed in predicting PPIs, assuming that homologous proteins exhibit similar interaction patterns and functions [31]. Traditionally, a pair of interacting proteins in one species and their corresponding orthologs in another species, known to interact with each other, are termed interaction-orthologs (interologs) [32, 33]. However, the concept can be expanded to include interaction-homologs, as the distinction between orthologs and paralogs is not always clear-cut [31, 34].

A similar concept to what we want to do was first proposed in 2001 by Chen and Zhi [35], in which it was called Inverse Docking. It refers to computationally docking a specific small molecule of interest to a library of receptor structures. The technique may be used to identify new potential biological targets of known compounds [36, 37, 38], or to identify targets for compounds among a family of related receptors [39]. Another application is to predict a compound's pharmacological profile [40] or to generate a virtual selectivity profile characterizing the inhibitors' promiscuity [41]. Given the multi-faceted nature of a pharmacologically active compound's biological effects, inverse docking is especially helpful, because it may generate new hypotheses for the action mechanism. In the case of inverse docking and similar techniques, to our knowledge, only small molecules are considered, whereas we include ligands/proteins without constraints on dimensions, sequence length or structure.

Deep learning models have the ability to learn patterns in data that are often unclear to us humans, making them difficult to spot using conventional methods. This explains why deep learning models are increasingly common in scientific research. For our purpose, the characteristics of deep learning models provide a clear advantage over standard bioinformatic tools and approaches. These models can identify proteins that are semantically relevant, which might be overlooked by other analyses and methods when studying interactions with a specific chosen protein. So choosing deep learning models to handle proteins information and representations, we think it would allow us to reach a higher probability of finding the most meaningful proteins for our purpose.

One notable deep learning technique is DSCRIPT, a sequence-based, structure-aware model for predicting protein-protein interactions. It bypasses the need for structural or experimental interaction data, capturing complex patterns within protein sequences indicative of interaction potential. It employs convolutional and recurrent neural networks to infer potential contact points between proteins, simulating the physical interaction space and generating an inter-protein contact map. DSCRIPT's key conceptual advance lies in implementing an interpretable, structure-based model despite having only sequence-based inputs. Leveraging recent advancements in protein language modeling [42], it constructs informative protein representations implicitly endowed with structural information. The model's generalizability and interpretability stem from its ability to learn informative geometric representations of proteins, transforming protein embeddings into a 2D contact map. The authors hypothesized that proteins with similar embeddings are likely to interact similarly, enabling the discovery of new interacting pairs. Evaluation against other protein sequence representations and BLAST [43] searches confirmed DSCRIPT's efficacy in identifying interactions. Additionally, the model's interpretability aids in predicting inter-protein docking contacts, producing contact maps consistent with the ground-truth contacts.

Given DSCRIPT's ability to generate meaningful protein representations and its interpretability through contact maps, it emerged as a logical choice for screening potential interaction protein candidates. Its inherent interpretability made it a preferred option for our approach.

To validate the proteins and interactions computed we use the SpeedPPI [7] model which represents a significant advancement in the field of computational biology, particularly in the rapid construction of protein-protein interaction networks. The model operates on the principle of evaluating pairwise interactions through AlphaFold2 following the FoldDock pipeline [44], which is specifically tailored for PPIs. We were inspired by this case study [45].

3. Methods

In this section, we present the mathematical foundation for our approach. We divide it into two parts: the former presents the first step of our approach, which consists of searching candidates and identifying proper filters; the latter focuses on validating the results.

3.1. Identification of potential candidates

The overall process starts from a known interaction between a target protein, T, and a ligand-protein, L. The aim is to find proteins that might interact with protein T similarly to L. We retrieve the proteins' data in the STRING [46] and UniProt [47] databases. After obtaining a set of proteins of interest, we pass them through the DSCRIPT model to get their representations.

Embeddings These representations are also referred to as "embeddings". An embedding is a numerical representation of objects, words, or entities in a continuous vector space. This representation usually encapsulates semantic and syntactic information in a dense, fixed-length vector format, facilitating computational operations.

DSCRIPT To retrieve the embeddings, we utilize a pre-trained model developed by Bepler and Berger [42], called DSCRIPT. This model consists of a Bi-LSTM [48] trained on three distinct types of data: the proteins' SCOP classification (i.e. Structural Classification of Proteins), which provides a general structural framework, the self-contact map detailing the protein's 3-D structure, and the sequence alignment among similar proteins. These embeddings effectively encapsulate the protein sequences' local context and overarching structural attributes. Specifically, the encoding for each amino acid is d_0-dimensional, capturing not only the properties of the individual amino acid but also the broader structural context of the entire protein sequence.

For a given protein L with sequence S_L of length n, DSCRIPT generates an embedding $E_AL \in \mathbb{R}^{n\times d_0}$, where d_0 is the dimensionality of the embedding space ($d_0 = 6165$ in our case). DSCRIPT's predictive capability hinges on its ability to generate embeddings for two proteins and compute an interaction probability $\hat{p} \in [0, 1]$. The model's architecture employs a projection module to reduce embeddings to a lower dimension d using a multi-layer perceptron using the rectified linear unit (ReLU). The projection module is followed by a residue-contact module, which takes the d-dimensional embeddings and models the interaction between the residues of each protein. This process results in a contact prediction matrix $\hat{C} \in [0, 1]^{n\times m}$ and a single interaction probability $\hat{p}$, derived from the matrix through global pooling operations, which captures the intuition that a pair of interacting proteins will be characterized by a relatively small number of high-probability interacting residues and a logistic activation function through the interaction prediction module. The global pooling operation captures the intuition that a pair of interacting proteins will be characterized by a relatively small number of high-probability interacting residues or regions This activation function takes the raw probability predictions and makes them steeper, depressing values below 0.5 towards 0 and inflating values above 0.5 towards 1, controlling the rate at which this occurs. Then $\hat{p}$ and $\hat{C}$ are returned as the model prediction.

Cosine Similarity In our approach, we use DSCRIPT first to obtain representations of all the proteins in the set, and then we compute a score that tells us how similar these representations are to the representation of the protein of interest. In particular, we leverage the cosine similarity:

$$cosine\_similarity(E_A, E_B) = \frac{\sum_{i=0}^{d} E_{Ai} E_{Bi}}{||E_A|| ||E_B||} \in [-1, 1], \quad (1)$$

where E_A and E_B are two embedding vectors and d is their dimension.

3.2. Validation of candidates

To validate the candidates, we use SpeedPPI, which is an optimized model based on AlphaFold2 [8] for PPI network prediction 40x faster and less disk space reliant [7].

Given an organism, the proteome is extracted from UniProt or another database. All sequences are used in single-sequence mode to create multiple sequence alignments (MSAs) with HHblits [49] searching on the Uniclust30 [50] database.

Multiple Sequence Alignments A multiple sequence alignment (MSA) organizes protein sequences into a rectangular array, aiming to align residues within columns that are homologous, superposable, or serve a common functional role, although these criteria may yield different alignments as sequence, structure, and function diverge over evolutionary time [51]. MSAs are indispensable in biology and bioinformatics for comparing sequences, revealing evolutionary relationships, and identifying conserved regions. They entail aligning nucleotide or amino acid sequences to detect similarities and differences, employing diverse algorithms, which aid in predicting functional domains and phylogenetic relationships but can face computational challenges with highly diverse sequences.

SpeedPPI and AlphaFold2 In the procedure, the MSAs are paired based on species information, and all single-chain information is maintained by block diagonalization. The MSA pairing and block diagonalization are done within the network, avoiding writing this information to disk, which reduces disk space requirements and the total prediction time. The structure prediction is made with AlphaFold2, and the features are prefetched. AlphaFold2 is a system designed to predict the 3D structures of proteins. It works by taking the sequence of amino acids that make up a protein and then using deep learning algorithms to predict how these amino acids fold and interact to form a 3D structure. Its architecture involves a deep neural network comprising two main components: a novel attention mechanism and a fully differentiable module for structure prediction. The attention mechanism allows the model to focus on the most relevant parts of the protein sequence when making predictions, while the structure prediction module uses a series of mathematical functions to estimate the distances between pairs of amino acids and the angles between connected amino acids, which are crucial for determining the overall shape of the protein. Additionally, AlphaFold2 uses a technique called “attention-based message passing” to incorporate information from the entire protein sequence into its predictions, allowing it to capture long-range interactions that were previously difficult to model accurately.

pDockQ score The predictions are evaluated with the pDockQ score within the prediction runs. The model produces pDockQ scores (introduced in [44, 45]). This score is created by fitting a sigmoidal curve, to the DockQ [52] scores:

$$pDockQ = \frac{L}{1 + e^{-k(x-x_0)}} + b$$

where x is the average interface plDDT multiplied with the logarithm of the number of interface contacts and L, x_0, k, b are obtained from the fitting process. From the prediction of the pDockQ score and 3D structure in AF2, we store the contact points of the interaction to validate that the interaction is similar to the one of the original ligand. The $pDockQ \geq 0.23$ was considered to indicate a well modeled interaction [45], hence we considered the protein used for the interaction to be a more probable ligand of NKp46. To further expand the analysis we generate a contact map and some histograms of the distribution of contact points in relation to NKp46 (see section 5).

4. Experiments

We start our experiments by validating the capability of PPI models to reproduce known interactions found in the literature involving NKp46. In particular, we observe that DSCRIPT recognizes the interactions of NKp46 with CFP and TYROBP. We focus our study mainly on the recently proved interaction CRT-NKp46 [9] which caught our interest, given its relevance in the medical field.

We generate high-dimensional representations of NKp46 and CRT through DSCRIPT. We compute the interaction score and contact map of the interaction (Figure 1).

From the analyses of the contact map, we notice that the most active amino acids of CRT in the interaction are in the range of 199-227, while for NKp46 they are in the range 80-258. Both ranges are exactly within the bounds of the CRT's P-domain and NKp46 extracellular region which are the regions involved in the interaction. This is coherent with what the authors of DSCRIPT claimed: DSCRIPT is capable of producing contact maps that are similar to the ground truth of the interaction [5].

We chose CRT as the template protein and we use the UniProt database[47] to find the initial set of proteins we are interested in. Focusing on human proteins, we filter the database accordingly and since for our case study the known interaction involves NKp46 and CRT, with CRT located on the cell surface/membrane after having translocated because of the cell ER stress, we refine our search to include only proteins likely to interact with NKp46 in this location. As a result, we narrowed down our database to 7188 proteins.

To find and filter the potentially similar proteins of CRT we try two approaches: (i) similarity to the entirety of the P-domain (amino acids 198-308) and (ii) similarity to the most small interacting range (triplet) of amino acids (199-201). There are two main reasons why we thought of these two approaches: (i) the P-domain is actively involved in the interaction [9] and finding proteins that have part of the amino acid sequence semantically similar to the entirety of the P-domain could lead to proteins with similar general characteristics to it, hence potential candidates for interaction; (ii) using the most "interacting" triplet of amino acids could be more

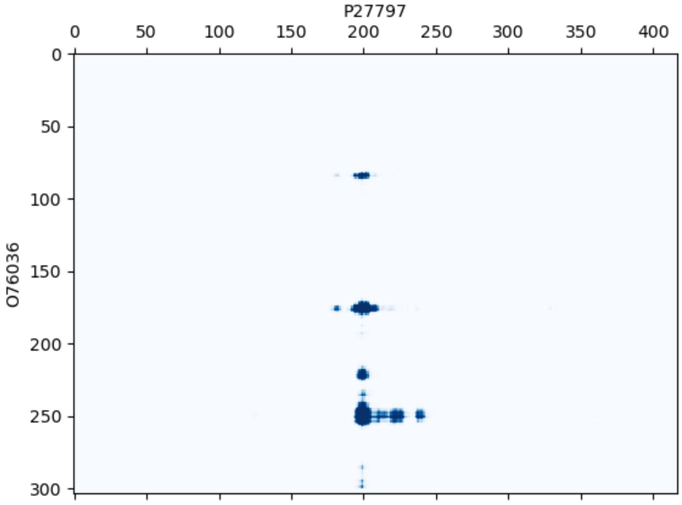


Figure 1: Contact map for the amino acids of NKp46 (on y axis) and CRT (on x axis).

precise because clearly the P-domain is not entirely interacting with NKp46 but most probably only a limited percentage of the sequence will be interested in the interaction, hence using the most interacting triplets could lead to a less noisy set of protein candidates compared to the other approach.

The high dimensional embeddings of the proteins are at the amino acid level so we compute the similarities at this exact level. In both cases, (i) and (ii), we use DSCRIPT to generate the proteins' representations. We analyze each protein in the database obtained, amino acid embedding per amino acid embedding of CRT. The similarity score at the amino acid level is given by the cosine similarity (see 3.1). In (i) we compare each amino acid representation of a protein to every amino acid of the P-domain in CRT, then we compute the mean and median. In (ii) we compare every triplet in the protein at hand and the most interacting triplet in CRT, then we compute the mean.

So a similarity score is assigned for both cases to each protein. Then, we pass such proteins to DSCRIPT in order to predict the interaction with NKp46, producing interaction scores and contact maps. The most important information is obtained: for (i), mean and median of the similarity to P-domain, DSCRIPT interaction score and contact map; for (ii) mean of the similarity between the most interacting triplets of the protein and CRT, interaction score and contact map. By combining this data we rank the proteins and obtain a limited list of candidates.

We take the best proteins overall (DSCRIPT interaction score ≥ 0.5 and highest median/mean similarity values) for the final validation through SpeedPPI. We compute the pDockQ score of the interaction and the 3D structure, so we can also get the predicted contact points. AlphaFold has some statistical components that may vary the outcomes in different executions, so we run the model 25 times with different seeds to obtain more robust final results.

We also run some experiments with commonly used bioinformatics tools such as BLAST [43] to find proteins similar to CRT. Hence we also used our approach on these proteins (and CRT) to compare the results.

In rest of the study we will refer to the proteins we found as Protein + index (P+index), to see the correspondence to the real proteins and their genes see table 3

5. Results

In this section, we analyze and compare the bioinformatics methodology (Section 5.1) with the two main directions taken for our experiments. As we introduced in Section 4 we consider the P-domain area (Section 5.2) and the three consecutive most interacting amino acids (Section 5.3) of CRT in order to find proteins similar to CRT that also have relevant features for interaction with NKp46.

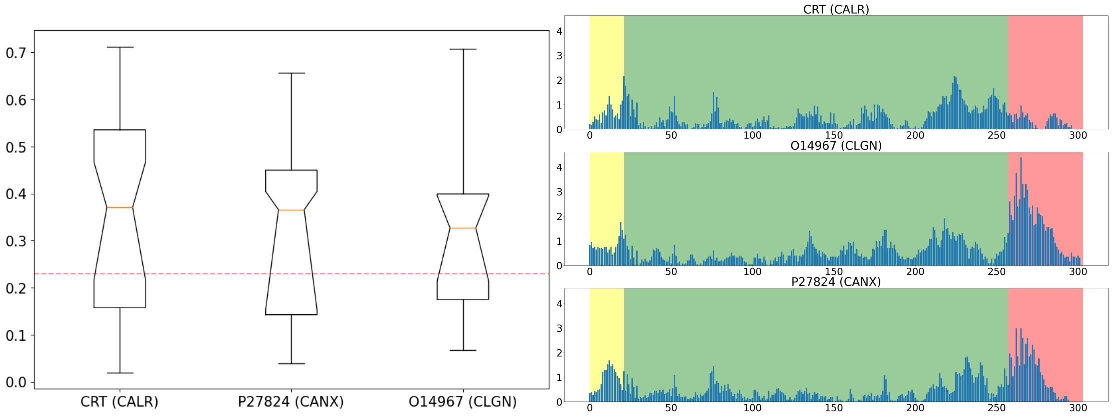
To analyze the obtained candidates we create the boxplot of pDockQ scores to better visualize the distribution of the results on the different runs of SpeedPPI that allow us to have robust outcomes. Then, knowing that the NKp46 is able to interact only in a specific portion of itself, the extracellular domain, we further investigate where the contacts effectively occur. For this reason, we report the distribution of the contact points along the NKp46 amino acids' sequence. We highlight the three main portions of NKp46 with different colors: yellow for the signaling peptide, green for the extracellular domain and red for the intracellular and transmembranal part. Additionally, we report in tables the average and median amount of contact points for area.

5.1. Bioinformatics methods

More commonly used tools that compute sequence and structure similarity between proteins, like BLAST [43], Prosite [53], MEME suite [54] or InterPro [55], when used to find similar proteins to CRT always found its isoforms or calnexin and calmegin. These latter proteins belong to the same protein family of CRT and have a similar sequence to the CRT's P-domain. Both of them reside in the endoplasmic reticulum (ER), share structural features and play essential roles in protein folding and quality control within the ER.

From Figure 2 and Table 1 we can see that while the pDockQ scores are high, most of the contacts are in the signaling (yellow) and extracellular (green) parts of NKp46 for CRT, instead for calmegin and calnexin the contacts tend to be mostly in the cytoplasmic part (red). It is reasonable that most contacts happen in the cytoplasmic part, due to the fact that they reside in the ER. CRT is also found in the ER, but when ER stress is induced it translocates to the membrane. We do not know if the same could happen to calmegin or calnexin, but the models seem to have captured this characteristic for CRT: the pDockQ score is high and the contacts are mostly on the extracellular part of NKp46 which is not the case for the other two proteins. This could mean that the similarity between calnexin, calmegin – which were found with bioinformatic tools – and CRT is more about sequence similarity than a similarity that alludes to a possibility for interaction.

Clearly, these bioinformatic techniques are precise and indeed return "similar" proteins in a sequence and structural sense but this kind of similarity is not enough, we think that using



(a) Boxplot of pDockQ scores

(b) Contact points distribution along NKp46 areas

Figure 2: Analysis made on calreticulin (CRT), calmegin (CLGN) and calnexin (CANX) when we predict their interaction with NKp46 using SpeedPPI model. In Figure (a) there are the pDockQ scores, while in (b) the contact points distribution on the different areas of NKp46. Both computed over 25 runs of the model.

		Calreticulin	**Calmegin**	**Calnexin**
Average	Signal	0.63	0.82	0.92
	Extra	0.51	0.52	0.43
	Intra	0.30	1.55	1.00
Median	Signal	0.56	0.76	0.96
	Extra	0.32	0.40	0.32
	Intra	0.28	1.56	0.64

Table 1
Predicted average and median contact frequencies by the SpeedPPI model across various regions of the NKp46 protein (i.e. signal peptide (Signal), Extracellular (Extra) and Transmembranal/Intracellular (Intra)) during interactions with calreticulin and candidates found via bioinformatics approach, namely calmegin and calnexin.

deep learning could lead us to a similarity related to the interacting ability that proteins may have thanks to its ability in learning hidden patterns.

5.2. P-domain Embedding Similarity

In experiment (i) we consider the P-domain area for computing cosine similarities and we extract the top-10 candidates, that we call P1...10. In Table 3 we report the UniProt IDs and genes that correspond to P1...10. In Figure 3 we can see the pDockQ scores obtained when predict the interaction between NKp46 and them.

Here we can see that only P1 has a pDockQ score ≥ 0.23. Figure 5, instead, shows the distribution of contacts for all the proteins with NKp46, here we notice that P1 has most of the contacts in the extracellular part with some high peaks. Hence, from the union of the information of both figures, we can consider P1 to be well modeled while not the other proteins

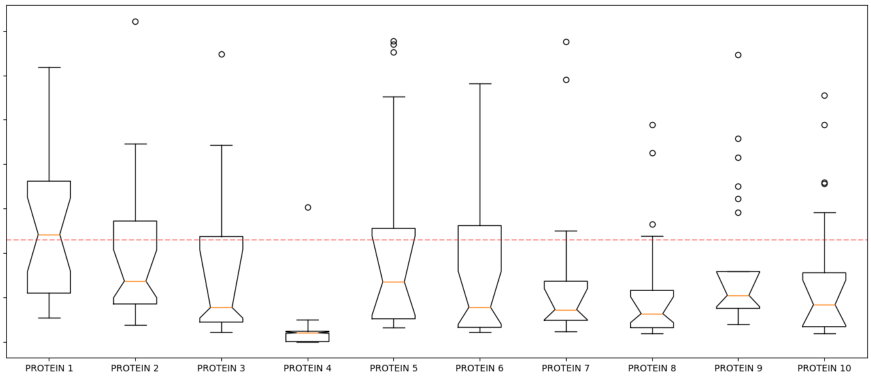


Figure 3: Boxplot of pDockQ values of the interaction between NKp46 and proteins having high similarity to the P-domain of CRT.

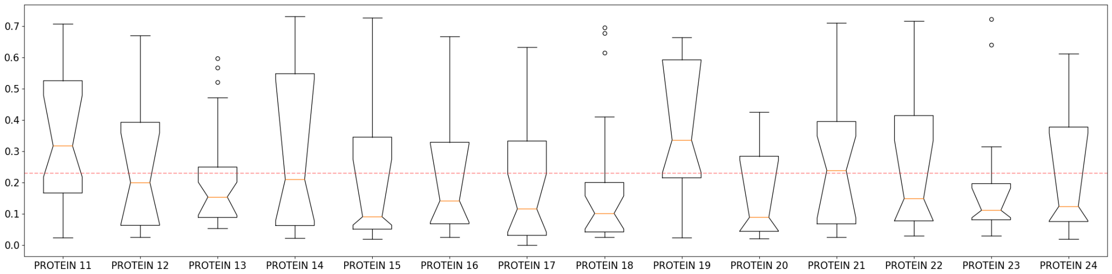


Figure 4: Boxplot of pDockQ values of the interaction between NKp46 and proteins amino acids' triplets having high similarity to the most interacting triplet of CRT.

that have contrasting data.

In Table 2a the average and median contact frequencies across the three regions we indicated (signaling, extracellular, intracellular regions) of the NKp46 protein are shown. From this data, we observe that the proteins obtained in the (i) P-domain similarity experiment have an overall low possibility of interaction. The data show slight inconsistencies between the pDockQ score computed and the distribution of frequencies of the contacts for each protein. For instance P2 has low pDockQ score and yet it has very high contacts' frequencies on the extracellular part; the same could be said of other proteins like P3.

In Figure 6 the mean and median of the number of contacts per amino acid over all the proteins are shown.

5.3. Triplet Embedding Similarity

In experiment (ii) with the triplet similarity we obtain protein P11-P24 (see Table 3 for their corresponding names and genes) and in Figure 4 we can see their computed pDockQ scores.

Here we can see that P11, P19 and P21 have a pDockQ score ≥ 0.23, hence we can consider

		P1	P2	P3	P4	P5	P6	P7	P8	P9	P10
Average	Signal	0.09	0.00	0.11	0.00	0.17	0.10	0.26	0.05	0.44	0.07
	Extra	0.26	0.33	0.31	0.04	0.30	0.26	0.15	0.10	0.27	0.15
	Intra	0.27	0.14	0.01	0.06	0.76	0.17	0.28	0.10	0.09	0.03
Median	Signal	0.00	0.00	0.00	0.00	0.12	0.08	0.24	0.00	0.20	0.04
	Extra	0.04	0.16	0.20	0.00	0.16	0.20	0.04	0.00	0.04	0.00
	Intra	0.24	0.04	0.00	0.00	0.48	0.08	0.20	0.20	0.00	0.00

(a) Considering P-domain area of CRT for computing similarities

		P11	P12	P13	P14	P15	P16	P17	P18	P19	P20	P21	P22	P23	P24
Average	Signal	0.89	0.45	0.39	0.39	0.16	0.19	0.22	0.06	1.40	0.04	0.00	0.50	0.39	0.60
	Extra	0.98	0.44	0.30	0.73	0.36	0.24	0.29	0.30	0.83	0.25	0.31	0.71	0.31	0.72
	Intra	0.32	0.26	0.31	0.45	0.13	0.63	0.25	0.05	1.20	0.40	0.25	0.60	0.05	0.26
Median	Signal	0.88	0.44	0.32	0.36	0.16	0.16	0.20	0.00	1.28	0.0	0.00	0.40	0.40	0.40
	Extra	0.84	0.32	0.20	0.52	0.24	0.12	0.28	0.20	0.08	0.08	0.08	0.52	0.16	0.68
	Intra	0.32	0.16	0.28	0.42	0.06	0.54	0.18	0.00	0.94	0.32	0.08	0.26	0.04	0.22

(b) Considering triplet of CRT's amino acids for computing similarities

Table 2
Predicted average and median contact frequencies by the SpeedPPI model across various regions of the NKp46 protein (i.e. signal peptide (Signal), Extracellular (Extra) and Transmembranal/Intracellular (Intra)) during interactions with the top-k proteins identified using our method. In Table (a) the results were obtained by using the P-domain (k=10) and in (b) by considering a triplet of aminoacids (k=14). The mapping between P1...24 and the protein names is shown in Table 3.

them to be well modeled with high probability while not the others are (note that P12 and P14 are near a pDockQ score of 0.23 so they could be decently modeled).

In Figure 7 the distribution of contact points over 25 run is shown. As in the (i) P-domain experiment, here in the (ii) triplets experiment, the coloured areas represent the sequence ranges 1-22 (yellow), 23-258 (green), 259-304 (red) and the green area is of most interest because it represents the extracellular part of NKp46. In this (ii) experiment it's clear that the quality of the proteins obtained is better, indeed we observe three proteins having high pDockQ scores and consistent distributions of contacts' frequencies with high values in the extracellular part and low values in the intracellular part. In Table 2b we report the average and median contacts' frequencies across the three regions we indicated (signaling, extracellular, intracellular) of the NKp46.

From this data, we observe that the proteins obtained in the (ii) triplet similarity experiment have better consistency. This difference in the results between the two experiments is related to the way we found the proteins. It seems that focusing on specific small regions (triplets) with the highest interacting score instead of large domains of interest to find proteins that share a high level of similarity can lead to more consistent results.

In Figure 8 the mean and median of the number of contacts per amino acid over all the proteins are shown.

In Figure 9 the mean and median of the number of contacts per amino acid over the high potential proteins P11, P19, P21 are shown.

6. Conclusions

In this work, we propose an approach to identify candidate proteins for interaction starting from a known interaction between a protein and its ligand, using such ligand as a template for the search. The method is based on combining the interpretability of the DSCRIPT model with the predictive power of AlphaFold.

We firstly checked the performance of the models on known interactions, and chose the recently discovered NKp46-CRT interaction as our case study. From UniProt we preliminarily filtered the proteins based on specific constraints: proteins found in the human body and specifically only on the membrane surface. We computed amino acids' high-dimensional representations of the proteins obtained from UniProt through the deep learning DSCRIPT model. We used DSCRIPT's inherent interpretability to make targeted choices on how to find similar proteins to CRT based on the NKp46-CRT interaction. We compared the results obtained with standard bioinformatics tools and our approach. Then we described the experiments: (i) filtering proteins based on a large domain of interest (CRT's P-domain in our case study); (ii) filtering proteins based on specific small most interacting region (triplet of amino acids) of the chosen protein (CRT in our case study). We used SpeedPPI (based on AlphaFold2) and the pDockQ score (from FoldDock) as a validation scheme.

Our findings reveal that this methodological approach can be useful in the search and analyses of PPIs when added to the more traditional bioinformatics tools typically employed for assessing proteins similarity. With our approach PPIs analyses can be more targeted and can help in finding new potential interacting proteins. We also introduce AlphaFold2 as a validation tool. The case study on NKp46's ligands demonstrates the method's ability to identify meaningful protein interactions, which could have implications for cancer immunotherapy.

In future, we would like to further investigate this case study with experimental validation from biological experts, as well as applying the approach to other proteins of interest.

Acknowledgments

We would like to express our deepest gratitude to Dr. Piera Fiore, Dr. Nicola Tumino and Dr. Paola Vacca from the Department of Immunology (IRCCS Ospedale Pediatrico Bambino Gesù, Rome, Italy) for their invaluable contributions and collaborative efforts in this study. Their expertise, guidance, and dedication were instrumental in advancing our research.

References

[1] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, et al., Highly accurate protein structure prediction with alphafold, Nature 596 (2021) 583–589.

[2] R. Evans, M. O'Neill, A. Pritzel, N. Antropova, A. Senior, T. Green, A. Žídek, R. Bates, S. Blackwell, J. Yim, et al., Protein complex prediction with alphafold-multimer, biorxiv (2021) 2021–10.

[3] P. Bryant, G. Pozzati, W. Zhu, A. Shenoy, P. Kundrotas, A. Elofsson, Predicting the structure of large protein complexes using alphafold and monte carlo tree search, Nature communications 13 (2022) 6028.

[4] C. Y. Lee, D. Hubrich, J. K. Varga, C. Schäfer, M. Welzel, E. Schumbera, M. Djokic, J. M. Strom, J. Schönfeld, J. L. Geist, et al., Systematic discovery of protein interaction interfaces using alphafold and experimental validation, Molecular Systems Biology 20 (2024) 75–97.

[5] S. Sledzieski, R. Singh, L. Cowen, B. Berger, Sequence-based prediction of protein-protein interactions: a structure-aware interpretable deep learning model, bioRxiv (2021). doi:10.1101/2021.01.22.427866.

[6] S. Sledzieski, R. Singh, L. Cowen, B. Berger, D-script translates genome to phenome with sequence-based, structure-aware, genome-scale predictions of protein-protein interactions, Cell Systems 12 (2021) 969–982.

[7] P. Bryant, F. Noé, Rapid protein-protein interaction network creation from multiple sequence alignments with deep learning, bioRxiv (2023). doi:10.1101/2023.04.15.536993.

[8] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, et al., Highly accurate protein structure prediction with alphafold, Nature 596 (2021) 583–589.

[9] S. Sen Santara, D.-J. Lee, A. Crespo, J. J. Hu, C. Walker, X. Ma, Y. Zhang, S. Chowdhury, K. F. Meza-Sosa, M. Lewandrowski, H. Zhang, M. Rowe, A. McClelland, H. Wu, C. Junqueira, J. Lieberman, The NK cell receptor NKp46 recognizes ecto-calreticulin on ER-stressed cells, Nature 616 (2023) 348–356. URL: https://www.nature.com/articles/s41586-023-05912-0. doi:10.1038/s41586-023-05912-0.

[10] T. Pazina, A. Shemesh, M. Brusilovsky, A. Porgador, K. S. Campbell, Regulation of the Functions of Natural Cytotoxicity Receptors by Interactions with Diverse Ligands and Alterations in Splice Variant Expression, Frontiers in Immunology 8 (2017). URL: https://www.frontiersin.org/articles/10.3389/fimmu.2017.00369.

[11] S.-Y. Wu, T. Fu, Y.-Z. Jiang, Z.-M. Shao, Natural killer cells in cancer biology and therapy, Molecular Cancer 19 (2020) 120. URL: https://doi.org/10.1186/s12943-020-01238-x. doi:10.1186/s12943-020-01238-x.

[12] N. K. Wolf, D. U. Kissiov, D. H. Raulet, Roles of natural killer cells in immunity to cancer, and applications to immunotherapy, Nature Reviews Immunology 23 (2023) 90–105. URL: https://www.nature.com/articles/s41577-022-00732-1. doi:10.1038/s41577-022-00732-1, number: 2 Publisher: Nature Publishing Group.

[13] O. Mandelboim, A. Porgador, NKp46, The International Journal of Biochemistry & Cell Biology 33 (2001) 1147–1150. URL: https://www.sciencedirect.com/science/article/pii/S1357272501000784. doi:10.1016/S1357-2725(01)00078-4.

[14] N. Bloushtain, U. Qimron, A. Bar-Ilan, O. Hershkovitz, R. Gazit, E. Fima, M. Korc, I. Vlodavsky, N. V. Bovin, A. Porgador, Membrane-associated heparan sulfate proteoglycans are involved in the recognition of cellular targets by NKp30 and NKp46, Journal of Immunology (Baltimore, Md.: 1950) 173 (2004) 2392–2401. doi:10.4049/jimmunol.173.4.2392.

[15] T. I. Arnon, H. Achdout, N. Lieberman, R. Gazit, T. Gonen-Gross, G. Katz, A. Bar-Ilan, N. Bloushtain, M. Lev, A. Joseph, et al., The mechanisms controlling the recognition of tumor-and virus-infected cells by nkp46, Blood 103 (2004) 664–672.

[16] J. R. Bock, D. A. Gough, Predicting protein–protein interactions from primary structure,

Bioinformatics 17 (2001) 455–460.

[17] E. Sprinzak, H. Margalit, Correlated sequence-signatures as markers of protein-protein interaction, Journal of molecular biology 311 (2001) 681–692.

[18] S. M. Gomez, W. S. Noble, A. Rzhetsky, Learning to predict protein–protein interactions from protein sequences, Bioinformatics 19 (2003) 1875–1881.

[19] A. Ben-Hur, W. S. Noble, Kernel methods for predicting protein–protein interactions, Bioinformatics 21 (2005) i38–i46.

[20] S. Martin, D. Roe, J.-L. Faulon, Predicting protein–protein interactions using signature products, Bioinformatics 21 (2005) 218–226.

[21] L. Nanni, A. Lumini, An ensemble of k-local hyperplanes for predicting protein–protein interactions, Bioinformatics 22 (2006) 1207–1210.

[22] S. Pitre, F. Dehne, A. Chan, J. Cheetham, A. Duong, A. Emili, M. Gebbia, J. Greenblatt, M. Jessulat, N. Krogan, et al., Pipe: a protein-protein interaction prediction engine based on the re-occurring short polypeptide sequences between known interacting protein pairs, BMC bioinformatics 7 (2006) 1–15.

[23] J. Shen, J. Zhang, X. Luo, W. Zhu, K. Yu, K. Chen, Y. Li, H. Jiang, Predicting protein–protein interactions based only on sequences information, Proceedings of the National Academy of Sciences 104 (2007) 4337–4341.

[24] Y. Guo, L. Yu, Z. Wen, M. Li, Using support vector machine combined with auto covariance to predict protein–protein interactions from protein sequences, Nucleic acids research 36 (2008) 3025–3030.

[25] S. Roy, D. Martinez, H. Platero, T. Lane, M. Werner-Washburne, Exploiting amino acid composition for predicting protein-protein interactions, PloS one 4 (2009) e7813.

[26] C.-Y. Yu, L.-C. Chou, D. T.-H. Chang, Predicting protein-protein interactions in unbalanced data using the primary structure of proteins, BMC bioinformatics 11 (2010) 1–10.

[27] J. Yu, M. Guo, C. J. Needham, Y. Huang, L. Cai, D. R. Westhead, Simple sequence-based kernels do not predict protein–protein interactions, Bioinformatics 26 (2010) 2610–2614.

[28] Y. Guo, M. Li, X. Pu, G. Li, X. Guang, W. Xiong, J. Li, Pred_ppi: a server for predicting protein-protein interactions based on sequence data with probability assignment, BMC research notes 3 (2010) 1–7.

[29] M. Deng, S. Mehta, F. Sun, T. Chen, Inferring domain-domain interactions from protein-protein interactions, in: Proceedings of the sixth annual international conference on Computational biology, 2002, pp. 117–126.

[30] M. Hayashida, M. Kamada, J. Song, T. Akutsu, Conditional random field approach to prediction of protein-protein interactions using domain information, BMC systems biology 5 (2011) 1–9.

[31] C.-C. Chen, C.-Y. Lin, Y.-S. Lo, J.-M. Yang, Ppisearch: a web server for searching homologous protein–protein interactions across multiple species, Nucleic acids research 37 (2009) W369–W375.

[32] L. R. Matthews, P. Vaglio, J. Reboul, H. Ge, B. P. Davis, J. Garrels, S. Vincent, M. Vidal, Identification of potential interaction networks using sequence-based searches for conserved protein-protein interactions or "interologs", Genome research 11 (2001) 2120–2126.

[33] H. Yu, N. M. Luscombe, H. X. Lu, X. Zhu, Y. Xia, J.-D. J. Han, N. Bertin, S. Chung, M. Vidal, M. Gerstein, Annotation transfer between genomes: protein–protein interologs and

protein–dna regulogs, Genome research 14 (2004) 1107–1118.

[34] J. Garcia-Garcia, S. Schleker, J. Klein-Seetharaman, B. Oliva, Bips: Biana interolog prediction server. a tool for protein–protein interaction inference, Nucleic acids research 40 (2012) W147–W151.

[35] Y. Chen, D. Zhi, Ligand–protein inverse docking and its potential use in the computer search of protein targets of a small molecule, Proteins: Structure, Function, and Bioinformatics 43 (2001) 217–226.

[36] Q.-T. Do, I. Renimel, P. Andre, C. Lugnier, C. D. Muller, P. Bernard, Reverse pharmacognosy: application of selnergy, a new tool for lead discovery. the example of ε-viniferin, Current drug discovery technologies 2 (2005) 161–167.

[37] P. Muller, G. Lena, E. Boilard, S. Bezzine, G. Lambeau, G. Guichard, D. Rognan, In s ilico-guided target identification of a scaffold-focused library: 1, 3, 5-triazepan-2, 6-diones as novel phospholipase a2 inhibitors, Journal of medicinal chemistry 49 (2006) 6768–6778.

[38] S. Zahler, S. Tietze, F. Totzke, M. Kubbutat, L. Meijer, A. M. Vollmar, J. Apostolakis, Inverse in silico screening for identification of kinase inhibitor targets, Chemistry & biology 14 (2007) 1207–1214.

[39] M. Schapira, R. Abagyan, M. Totrov, Nuclear hormone receptor targeted virtual screening, Journal of medicinal chemistry 46 (2003) 3045–3059.

[40] J. M. Rollinger, Accessing target information by virtual parallel screening—the impact on natural product research, Phytochemistry Letters 2 (2009) 53–58.

[41] C. Bissantz, A. Logean, D. Rognan, High-throughput modeling of human g-protein coupled receptors: amino acid sequence alignment, three-dimensional model building, and receptor library screening, Journal of chemical information and computer sciences 44 (2004) 1162–1176.

[42] T. Bepler, B. Berger, Learning protein sequence embeddings using information from structure, CoRR abs/1902.08661 (2019). URL: http://arxiv.org/abs/1902.08661. arXiv:1902.08661.

[43] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, D. J. Lipman, Basic local alignment search tool, Journal of Molecular Biology 215 (1990) 403–410. URL: https://www.sciencedirect.com/science/article/pii/S0022283605803602. doi:https://doi.org/10.1016/S0022-2836(05)80360-2.

[44] P. Bryant, G. Pozzati, A. Elofsson, Improved prediction of protein-protein interactions using alphafold2 and extended multiple-sequence alignments, bioRxiv (2021). doi:10.1101/2021.09.15.460468.

[45] D. F. Burke, P. Bryant, I. Barrio-Hernandez, D. Memon, G. Pozzati, A. Shenoy, W. Zhu, A. S. Dunham, P. Albanese, A. Keller, R. A. Scheltema, J. E. Bruce, A. Leitner, P. Kundrotas, P. Beltrao, A. Elofsson, Towards a structurally resolved human protein interaction network, bioRxiv (2021). doi:10.1101/2021.11.08.467664.

[46] D. Szklarczyk, R. Kirsch, M. Koutrouli, K. Nastou, F. Mehryary, R. Hachilif, A. L. Gable, T. Fang, N. T. Doncheva, S. Pyysalo, et al., The string database in 2023: protein–protein association networks and functional enrichment analyses for any sequenced genome of interest, Nucleic acids research 51 (2023) D638–D646.

[47] T. U. Consortium, UniProt: the Universal Protein Knowledgebase in 2023, Nucleic Acids Research 51 (2022) D523–D531. URL: https://doi.org/10.1093/nar/gkac1052.

doi:10.1093/nar/gkac1052. arXiv:https://academic.oup.com/nar/article-pdf/51/D1/D523/48441158/gkac1052.pdf.

[48] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural computation 9 (1997) 1735–1780.

[49] M. Remmert, A. Biegert, A. Hauser, J. Söding, Hhblits: lightning-fast iterative protein sequence searching by hmm-hmm alignment, Nature methods 9 (2012) 173–175.

[50] M. Mirdita, L. von den Driesch, C. Galiez, M. J. Martin, J. Söding, M. Steinegger, Uniclust databases of clustered and deeply annotated protein sequences and alignments, Nucleic Acids Research 45 (2016) D170–D176. URL: https://doi.org/10.1093/nar/gkw1081. doi:10.1093/nar/gkw1081. arXiv:https://academic.oup.com/nar/article-pdf/45/D1/D170/8846789/gkw1081.pdf.

[51] R. C. Edgar, S. Batzoglou, Multiple sequence alignment, Current Opinion in Structural Biology 16 (2006) 368–373. URL: https://www.sciencedirect.com/science/article/pii/S0959440X06000704. doi:https://doi.org/10.1016/j.sbi.2006.04.004, nucleic acids/Sequences and topology.

[52] S. Basu, B. Wallner, Dockq: a quality measure for protein-protein docking models, PloS one 11 (2016) e0161879.

[53] C. J. A. Sigrist, E. de Castro, L. Cerutti, B. A. Cuche, N. Hulo, A. Bridge, L. Bougueleret, I. Xenarios, New and continuing developments at PROSITE, Nucleic Acids Research 41 (2012) D344–D347. URL: https://doi.org/10.1093/nar/gks1067. doi:10.1093/nar/gks1067. arXiv:https://academic.oup.com/nar/article-pdf/41/D1/D344/3608670/gks1067.pdf.

[54] T. L. Bailey, J. Johnson, C. E. Grant, W. S. Noble, The MEME Suite, Nucleic Acids Research 43 (2015) W39–W49. URL: https://doi.org/10.1093/nar/gkv416. doi:10.1093/nar/gkv416. arXiv:https://academic.oup.com/nar/article-pdf/43/W1/W39/17435890/gkv416.pdf.

[55] S. Hunter, R. Apweiler, T. K. Attwood, A. Bairoch, A. Bateman, D. Binns, P. Bork, U. Das, L. Daugherty, L. Duquenne, R. D. Finn, J. Gough, D. Haft, N. Hulo, D. Kahn, E. Kelly, A. Laugraud, I. Letunic, D. Lonsdale, R. Lopez, M. Madera, J. Maslen, C. McAnulla, J. McDowall, J. Mistry, A. Mitchell, N. Mulder, D. Natale, C. Orengo, A. F. Quinn, J. D. Selengut, C. J. A. Sigrist, M. Thimma, P. D. Thomas, F. Valentin, D. Wilson, C. H. Wu, C. Yeats, InterPro: the integrative protein signature database, Nucleic Acids Research 37 (2008) D211–D215. URL: https://doi.org/10.1093/nar/gkn785. doi:10.1093/nar/gkn785. arXiv:https://academic.oup.com/nar/article-pdf/37/suppl_1/D211/3287888/gkn785.pdf.

A. Experiments details

In Table 3 there is the mapping from P1...24 proteins and the protein ID in UniProt database and the corresponding gene name.

Name in paper	Protein ID	Gene
P1	P15328	FOLR1
P2	P14207	FOLR2
P3	H0Y6X0	TAS1R1
P4	B7Z487	-
P5	O00451	GFRA2
P6	P09603	CSF1
P7	P56159	GFRA1
P8	Q6ULR6	-
P9	O60609	GFRA3
P10	Q9P0L0	VAPA
P11	A0A7I2V2B1	NRXN3
P12	A0A0D9SEM5	NRXN1
P13	A4FVB9	NRXN1
P14	A0A0A0MTQ4	RGMA
P15	A0A1B0GTB0	ATP6AP2
P16	A0A0H3VB22	FASLG
P17	A0A0U1RR22	PACSIN2
P18	A0A8I5KWD3	AP2M1
P19	A0A0U1RRJ0	NRXN3
P20	A0A977WMN2	TNF
P21	A6ND01	IZUMO1R
P22	O00481	BTN3A1
P23	P58400	NRXN1
P24	Q9HDB5	NRXN3

Table 3
Mapping between P1...24 and the corresponding protein ID in UniProt and their gene name if available.

B. Results details

In this section, we report some plots that show the distribution of the contact points' frequencies in NKp46 protein when interacting with other proteins. In Figure 5 there are the contact points of the top-10 candidates obtained through our method when considering the similarity of potential ligands with the P-domain of CRT. Additionally, in Figure 6, the average and median frequencies of the contact points for the same experiment are plotted. We discuss the observations about these results in Section 5.2.

Similarly, we report the same plots but for the experiments considering the most interacting amino acids' triplet of CRT for computing the similarity scores with proteins. So, in Figure 7 there are the distributions for each of the top-14 candidates, while in Figure 8 the average and median on such candidates. We discuss these results in Section 5.3.

Lastly, in Figure 9 there are the plots of the frequency distributions of the contact points for the candidates for which the SpeedPPI model predicts the interaction. In particular, the upper image reports the average of the contact points along NKp46 amino acids divided by areas and the bottom one the median. We also review these results in Section 5.3.

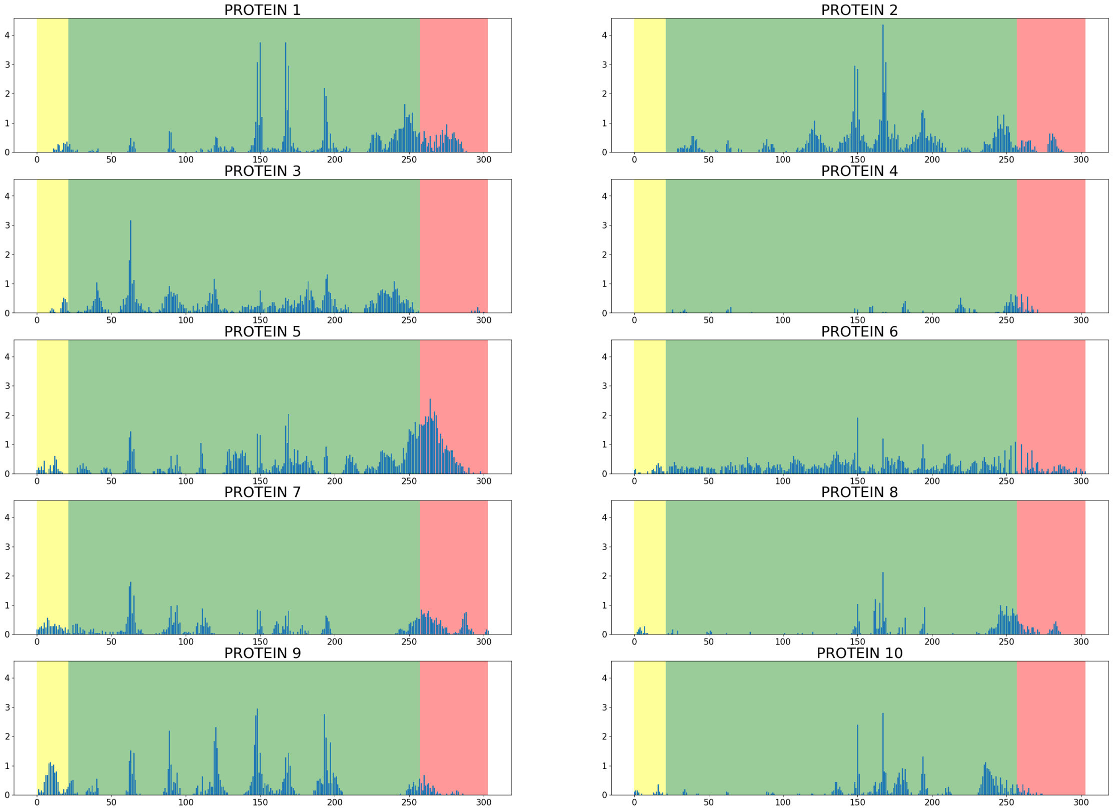


Figure 5: Distribution of the number of contact points for each protein in the (i) P-domain experiment.

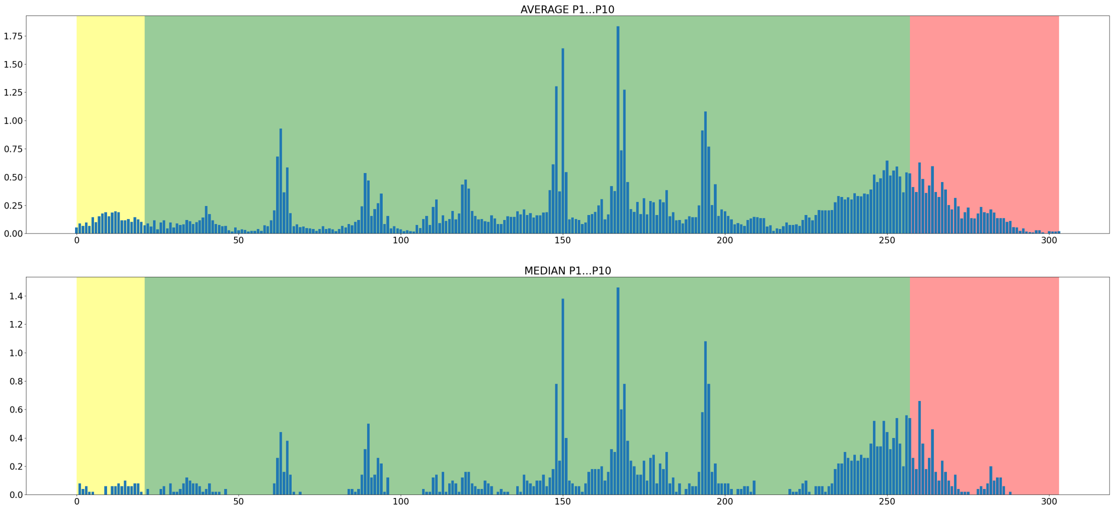


Figure 6: The mean and median of the number of contacts per amino acid over all the proteins in the (i) P-domain experiment.

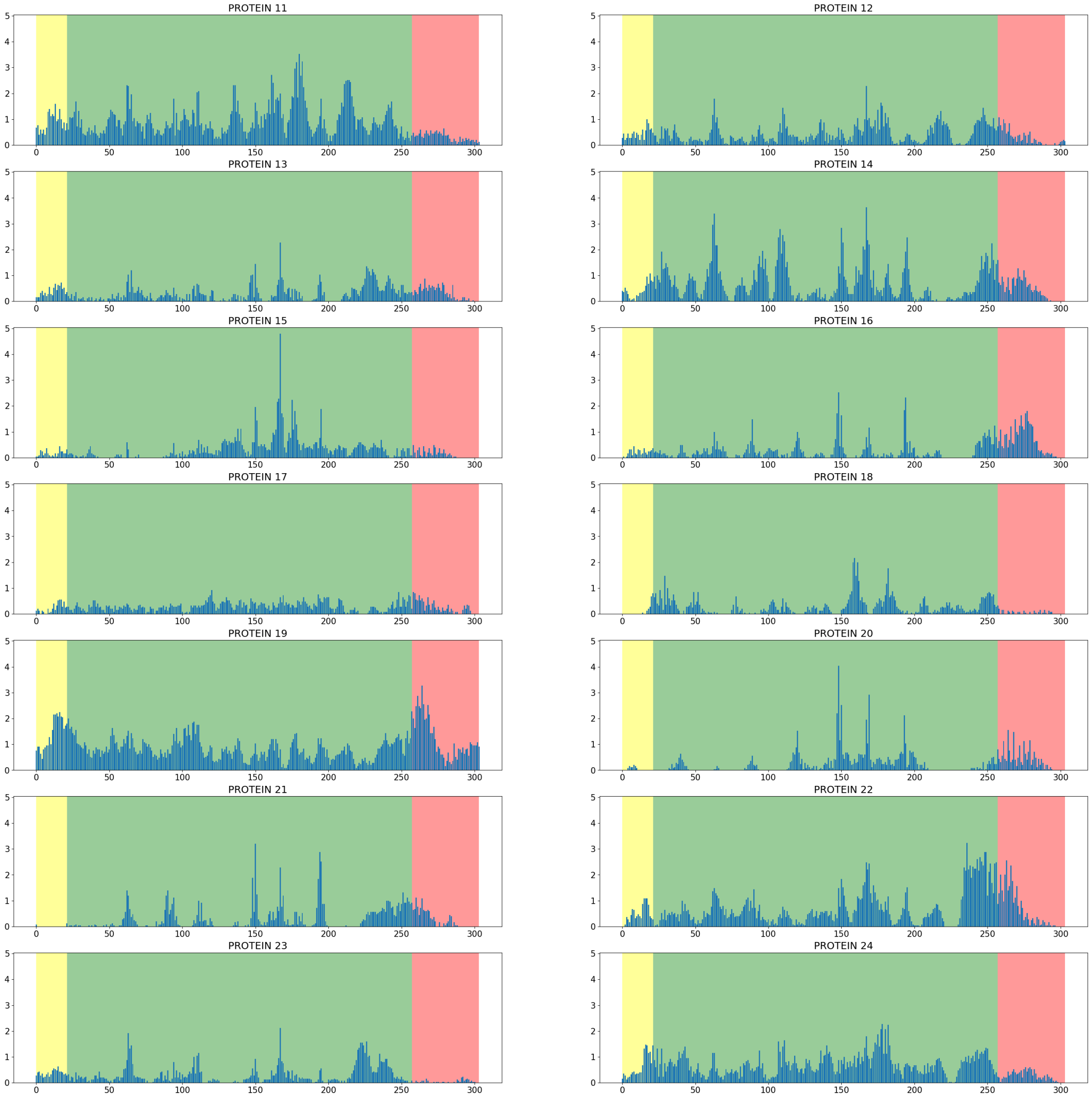


Figure 7: Distribution of the number of contact points for each protein in the (ii) triplets experiment.

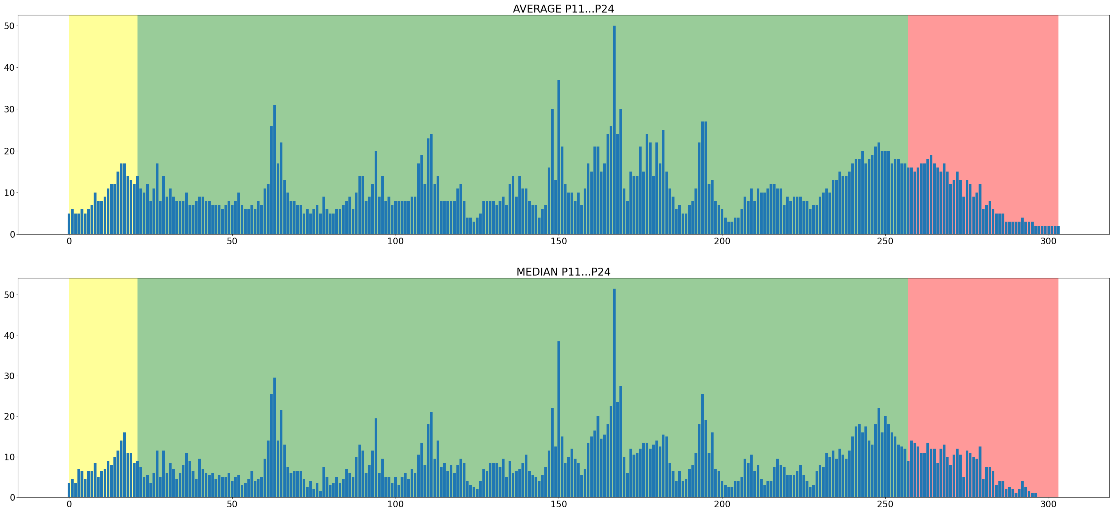


Figure 8: The mean and median of the number of contacts per amino acid over all the proteins in the (ii) triplets experiment.

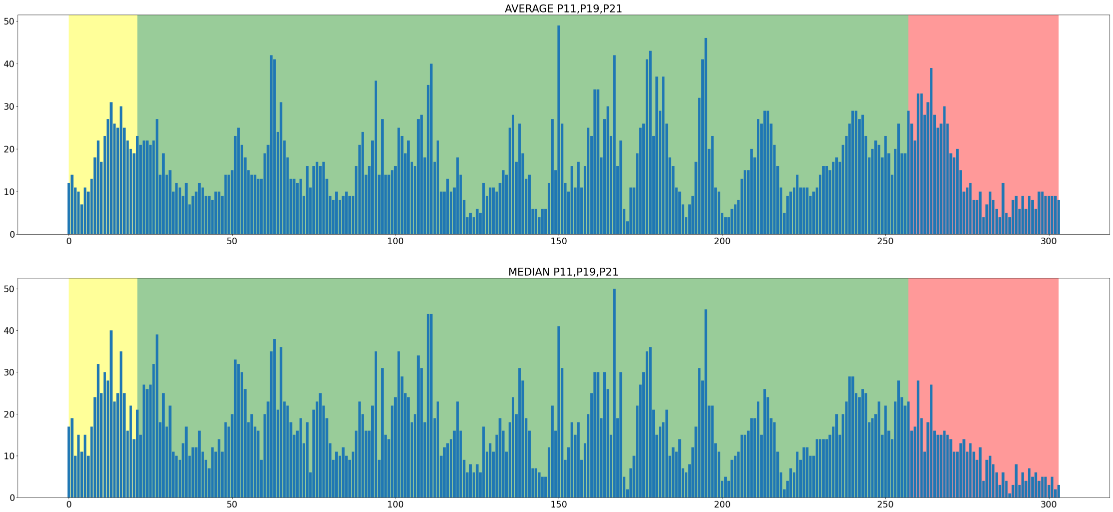


Figure 9: The mean and median of the number of contacts per amino acid over the high potential proteins P11, P19, P21 in the (ii) triplets experiment.