UNIVERSITAT DE BARCELONA

# Protein-protein docking for interactomic studies and its aplication to personalized medicine

Didier Barradas Bautista

# UNIVERSITAT DE BARCELONA

Facultat de Farmàcia


Programa de Doctorat en Biomedicina

RD 99/2011

# Protein-protein docking for interactomic studies and its aplication to personalized medicine

Memòria presentada per Didier Barradas Bautista

per optar al títol de doctor per la Universitat de Barcelona


Director

Dr Juan Fernández-Recio

Tutor

Dr Josep Lluis Gelpí Buchaca

Doctorand

Didier Barradas Bautista

Barcelona Supercomputing Center



UNIVERSITAT DE BARCELONA

# Index

# Figure index

## Table index



## Suplementary figures index

# Supplementary table index

# Acknowlegments

In the first place I would like to thanks my advisor Juan Fernandez-Recio for the opportunity to study my PhD degree in Europe. During my stay here at the Barcelona supercomputing center I have develop my skill in computational biology and more. It could not be possible if Juan in the first place did not submitted a call with the CONACyT. Therefore, thank you again Juan.

Also thanks to CONACyT for the funding of the PhD call for the BSC. And, thanks to BSC itself, for having me as part of the scientific staff, and for allow me to use a Marenostrum and therefore giving me the coolest scientific presentation line: “I work with a supercomputer, you know?”

I'd also want to thank Dr Humberto Lanz, Dr Mario Henri Rodriguez, and specially Dr Salvador Hernandez, all of them lend me a hand with the recommendation letter because they know me very well from my bachelor and master degree but it was Dr “Chava” who started me in this path and taught me about it. So thank you all three for the support.

Is a most, to thank all the guys and girls in the office and accessory people, this journey is not the same without companions. Thank you very munch to Israel, Sandra, Fatima, Montse, Santi , Armin, Ryoyi, Marina, Agusti, Jelissa, Ferran, Gerard, Marce, Silvia, Txema, Valenti, Leyden, Pedro, Jorge (From Zaragoza), Dimitry (who make me believe that Russian are the god guys in history), Laia, Oscar, Nacho Viciano and Ignacio Soteras, Rose, Chivis, Ciara, Freddy, Gabs(Gabriela),Romina (for the super healthy pastries), and all those that I don’t remember now. A very special thanks to Emanuele Monza, buddy without you and your Italianess nothing wouldn't be so funny and of course nothing wouldn't be so interesting to talk, thank you my friend.

To the PPD team, my good friends : Iain (publishing machine fueled by beer), Mark (Markimus), Dhoha, Miguel, Laura ,Chiara, Sergio, Luis (the boss killer), Jorge, of course “Brayan” (Python guru powered by Coca-cola), Mirella and Lucia (Antonia's team). I experienced so many things these years with all of you and all that I take it to heart and forever in my memory.

Now I have not forgotten my beloved Mexican friends. Ailett, Arturo, Carlos and Beto (¡¿Betty?!). As it seem it matters not where and how far I travel, how old, crazy or reluctant I become, these four find the way to find me. I see a bit of me in all you, that's way I worry and that's why I use social media to stay constantly in touch. Ailett and Arturo have been with me since a long time ago, you know and I thank for that. Carlos and Beto have been close to me since bachelor and could not be more grateful for the challenge that you two crazy aneuploids present, ever since I met you

I know that I will have a blast just trying to outsmart you. Thank you all.

Of course, to my wife **Mariana**, always supporting me, always loving me, love you back baby girl and thank you or being this strong for me and for us. To my son, because he gave strength every day I crossed the city in my old bike, every day that I complained, he became the will to continue, you my son are very special, handsome, intelligent talented, I mean of course he is I made half of him. Iker, you probably never read this thesis, by the time you became a skilled science reader probably this will be outdated but be sure this work was fueled, and inspired by you and your mom.

A la familia Lopez Serena, muchas gracias por el apoyo y el amor compartido en estos años apesar de la diststancia, estoy muy feliz de conocerlos y vivir con ustedes parte de esta aventura.

A mi familia, mama y papa, me duele no haber podido verlos en años, ni haber estdo alli, pero siempre llevo conmigo el trabajar duro y me alivia sabe rque aprueban que “asi era la jugada” a pesar de la distancia, a mis herman@s (Kevin, Zaid, Marya y Michelle ), como los he extrañado todos los dias, en estos años que no pude regresar a verlos me he perdido de fiestas, de cumpleaños y seguro de mas momentos, pero se que aun me quieren (tampoco es que tengan otro hermano mayor de reemplazo), los quiero enan@s, saben que siempre sirven de inspiracion y fuerza.

I finally will become a PhD, I was sure, and now I am scared and exited, as many of my fellows from the scientific community the future seems to be uncertain as our high specialization is not look as an asset, maybe more like a burden. For me to land a job is going to be hard, as many of my fellows, but anyway is has never been easy for me. So as I see the exit for my student status, also I see the rise of the challenge, I will be up to it.

As this chapter of my life closes, Barcelona has a very special place in my heart and mind, This city truly is marvelous, with things to do as family or as single, as host or visitor, Barcelona is munch more than a soccer team, I know is the most famous asset of the city, but beyond that there is so munch more, the science, the culture of respect, the gastronomy and the health culture. Is so easy to get comfortable here, that it strike me that there is people out there that do not know this. Maybe is because of the madness in my own country that I find this fascinating.

So, now read on, this thesis is the compendium of my four year journey, please be kind, and I hope you find it interesting.

Didier

# Summary

Proteins are the embodiment of the message encoded in the genes and they act as the building blocks and effector part of the cell. From gene regulation to cell signalling, as well as cell recognition and movement, protein-protein interactions (PPIs) drive many important cellular events by forming intricate interaction networks. The number of all non-redundant human binary interactions, forming the so-called interactome, ranges from 130,000 to 650,000 interactions as estimated by different studies. In some diseases, like cancer, these PPIs are altered by the presence of mutations in individual proteins, which can change the interaction networks of the cell resulting in a pathological state. In order to fully characterize the effect of a pathological mutation and have useful information for prediction purposes, it is important first to identify whether the mutation is located at a protein-binding interface, and second to understand the effect on the binding affinity of the affected interaction/s. To understand how these mutations can alter the PPIs, we need to look at the three-dimensional structure of the protein complexes at the atomic level. However, there are available structures for less than 10% of the estimated human interactome. Computational approaches such as protein-protein docking can help to extend the structural coverage of known PPIs.

In the protein-protein docking field, rigid-body docking is a widely used docking approach, since is fast, computationally cheap and is often capable of generating a pool of models within which a near-native structure can be found. These models need to be scored in order to select the acceptable ones from the set of poses. In the present thesis, we have characterized the synergy between combination of protein-protein docking methods and several scoring functions. Our findings provide guides for the use of the most efficient scoring function for each docking method, as well as instruct future scoring functions development efforts

Then we used docking calculations to predict interaction hotspots, i.e. residues that contribute the most to the binding energy, and interface patches by including neighbour residues to the predictions. We developed and validated a method, based in the Normalize Interface Propensity (NIP) score.

Summary

The work of this thesis have extended the original NIP method to predict the location of disease-associated nsSNPs at protein-protein interfaces, when there is no available structure for the protein-protein complex. We have applied this approach to the pathological interaction networks of six diseases with low structural data on PPIs. This approach can almost double the number of nsSNPs that can be characterized and identify edgetic effects in many nsSNPs that were previously unknown. This methodology was also applied to predict the location of 14,551 nsSNPs in 4,254 proteins, for more than 12,000 interactions without 3D structure. We found that 34% of the disease-associated nsSNPs were located at a protein-protein interface. This opens future opportunities for the high-throughput characterization of pathological mutations at the atomic level resolution, and can help to design novel therapeutic strategies to re-stabilize the affected PPIs by disease-associated nsSNPs.

**Keywords:** Single Nucleotide Polymorphisms (SNPs), pathological mutations, protein-protein docking, binding hot-spot, interface predictions, disease pathways

"I have wrestled with an alligator,
I've done tussled with a whale,
I've  handcuffed lightnin',
thrown thunder in jail.
Only last week, I murdered a rock,
injured a stone, hospitalized a brick.
I'm so mean, I make medicine sick.
All you chumps are gonna bow
when I whoop him, all of you.
I know you got him,
I know you've got him picked,
but the man is in trouble.
**I'm gonna show you how great I am.**"
- Muhammed Ali

# Chapter 1 Introduction

## 1.1 The role of the proteins in the cell

A cell is the basic structural and functional unit of any living organism. From single cell organisms to multicellular organisms, most of the cells have information stored in the DNA, coded in the form of nucleotide sequences, which must be transcribed into RNA, and then in turn into a chain of amino acids, the building blocks of proteins. This straightforward flux of information is the so-called “central dogma” (Crick 1970). However, this linear view of the flow of information is incomplete. In nature, self-interacting elements capable of modifying the above described flux of information challenge the idea of the central dogma. This is the case of ribozymes with self-catalytic activity (Lilley and Fritz, n.d.), and prions (Derkatch and Liebman 2007), misfolded proteins that can alter the structure and function of other proteins. These self-interacting elements add loops to the straight line in the central dogma. Even with these added loops, this view does not fully depict the crowded and dynamic environment inside the cell. There are additional genetic mechanisms that regulate the levels of proteins. An example of this is the field of epigenetics where the marks found in the DNA nucleosomes, such as methylation, prevents the transcription of DNA (Bharathy and Reshma 2012). Proteins themselves appear to have an active role to protect the balance of gene products even when the cell present an abnormal load of the genetic material like in polyploidy (Stingele et al. 2012). Among all of the interactions and factors that are driving all these processes, proteins have a prominent role as they can serve as scaffolds, provide protection to RNA or DNA (chaperones and nucleosomes), and act as receptors or effectors (such neuropeptides and enzymes).

Most proteins do not act as isolated units, and their interactions with biomolecules including other proteins are a fundamental property that gives rise to different cellular events (Stingele et al. 2012; Teichmann 2002). A fundamental aspect that should be taken into account is the three-dimensional (3D) nature of cellular components. Schematic representation of DNA, RNA and proteins have been two-dimensional for simplicity, but all three folds into three-dimensional space.

Protein folding is an exciting process itself. From the amino acid sequence, the inherent physicochemical properties of the polypeptide chain determine the first level of folding, known as

secondary structure, which can be either **β**-sheets or **α**-helices, as well as elements of the sequence that do not fold into a specific structure and are referred to as “loops”. From this, combinations of **β**-sheets and **α**-helices can form the tertiary structure, where many proteins gain their functionality. However, the majority of cell processes require the assembly of protein complexes, which constitute the so-called quaternary structure.

The relationship between the genetic information contained in the DNA and the structure of proteins is currently object of intense investigation. Recent sequencing efforts have yielded much information on the variants in genes (mutations), and association studies have revealed that these variations are tightly linked to the physiological outcome of the organism (Lander 2011; Freedman et al. 2011). There are two major approaches to analyze the effect of these variants: a reductionist view where the analysis is focused on the molecular effect of a mutation based on the 3D atomic structure of the protein of interest, and a systems approach focused on the effect on the network generated by the interactions between the elements in the cell (**Figure 1**). The present thesis shows that the synergy between these two approaches provides understanding on how mutations can affect interactions from an atomic level to the network organization.

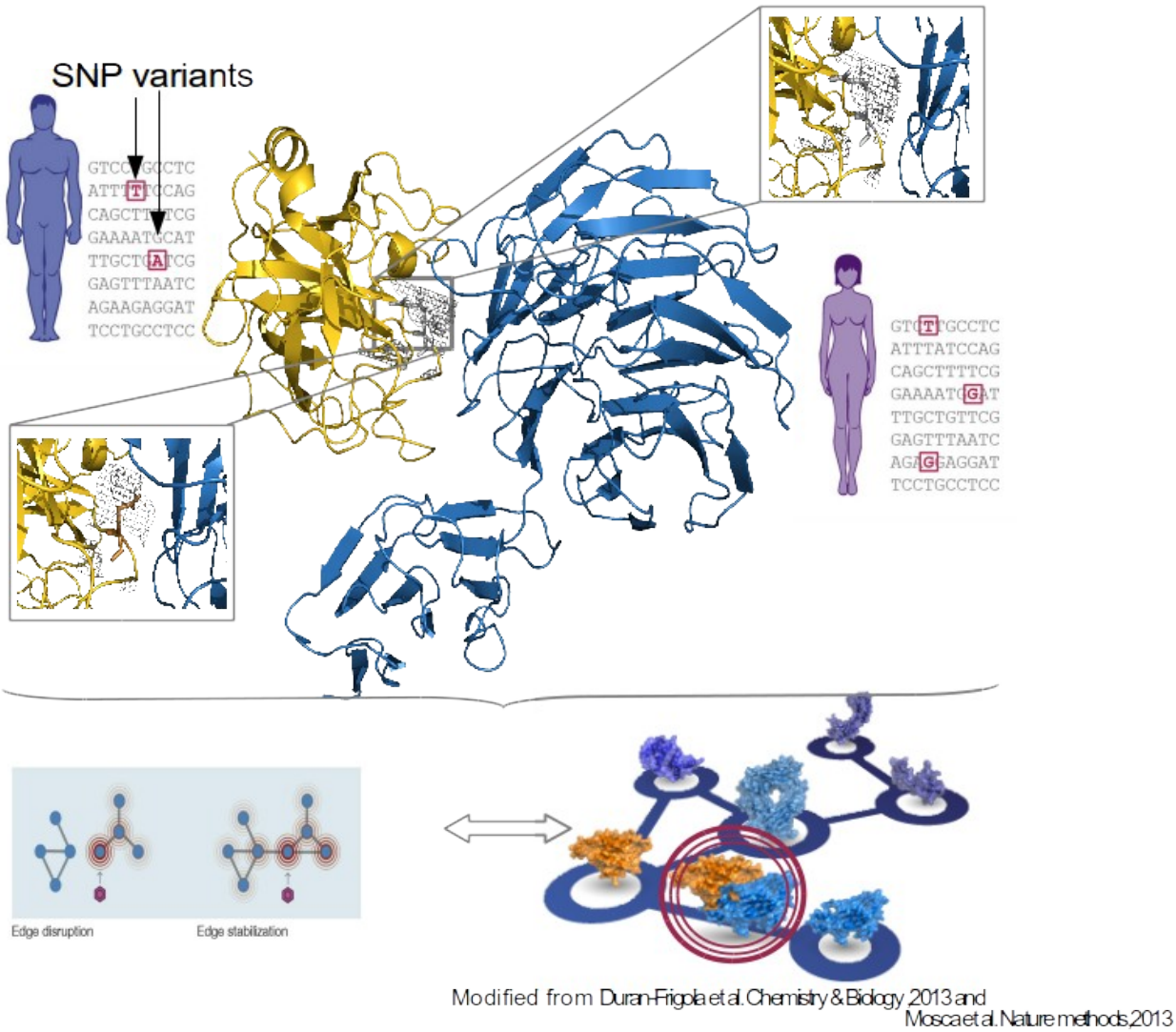



*Figure 1:* ***Influence of the nsSNPs in the structure and the PPI network.***

The appearance of a nsSNPs in the human genome can modify the structure of the protein. A small group of nsSNPs alter the interface of proteins. This can have an effect at the network scale in the cell

## 1.2 The road to personalized medicine

High-throughput techniques, like genome sequencing, mass spectroscopy and DNA and RNA expression microarrays, are dramatically changing the way we study biological sciences. The first major change arises from the massive data generated by these techniques. Next-generation sequencing (NGS) technologies have dramatically lowered the costs of gene sequencing, and are providing genomic information for an increasing number of healthy individuals and patient populations. A biological scientist has to face the overwhelming stream of information from

different sources, ranging from microorganisms (Venter et al. 2004) to patients in health care systems (Wang 2014). Computational resources are fundamental to efficiently analyze all this data. Institutes like National Center for Biotechnology (NCBI) and the European Bioinformatics Institute (EBI) receive data from different sources and store it in big public databases such the GenBank (Benson et al. 2005) and UniProt (The UniProt Consortium 2007). Moreover, they have integrated a variety of tools like BLAST (Benson et al. 2005) or CLUSTAL (Higgins and Sharp 1988) in publicly available websites with the goal of providing the scientific community with analytical tools for their research. This vast amount of information is an opportunity for biological sciences to put statistical methods and rigorous mathematical models into the molecular details that rule a living organism.

Following the first human genome completion (International Human Genome Sequencing Consortium 2004), the scientific community started an international effort known as the “1000 genomes project” (1000 Genomes Project Consortium et al. 2015). The project, now finished, consisted in obtaining the genome sequence from subpopulation around the world, making the genomes available to the scientific community for a variety of analysis. It also provides a framework for important questions on human genetics. In the past, the study of the genetic variation in the human population or genotypes was only possible using the unique gene variants that gave rise to evident distinct states or phenotypes. While the term genotype refers to the information stored in the DNA sequences, the term phenotype refers to the product of the genotype or “what we can see”, which can mean a protein fold or a cell type or even the look of the whole organism. With the lower costs of genome sequences and resources like the "1000 genomes", common genetic traits were found to be present in a large proportion of the human population (International HapMap 3 Consortium et al. 2010). Many of these traits were determined by **Single Nucleotide Polymorphisms (SNPs)**. SNPs are single base pair changes in the DNA sequence that occur with high frequency on the human genome (1000 Genomes Project Consortium et al. 2010) and the field of human genetics now use it as the unit for genetic variation in populations. When the change occur in the coding regions of DNA and results in a change in the amino acid of the protein is called non-synonimous SNPs (**nsSNPs**). The nsSNPs can either produce a nonfunctional protein (nonsense substitutions) or a multifunctional protein (missense substitutions) and in both cases can leading to a disease phenotype (Al-Haggar et al. 2012; Cordovado et al. 2012). The International HapMap Project aims to identify changes among the genomes and to find correlations with the observed

phenotypes. The number of SNPs per human genome is estimated to be around 10 million, all of them showing a different effect. HapMap has so far catalogued 1.6 million SNPs with genotypes from 11 human populations, including Japanese population from Tokyo, the Yoruba population from Africa, Han Chinese from Beijing, and European descent population (International HapMap Consortium 2005; Ritchie et al. 2010; International HapMap 3 Consortium et al. 2010).

**Genome-Wide Association Studies** (**GWAS**) are a powerful tool to identify a link of a relevant SNP with a human disease (Welter et al. 2013). The goal of GWAS is to identify genetic risk factors through various association tests, backed by statistical analysis, to make predictions about who is predisposed to a given disease, and then determine the genetic interplay of disease susceptibility for the development of new therapeutic strategies (Bush and Moore 2012). The most successful application of GWAS has been the identification of DNA sequences that play a role in drug response (metabolism, efficacy or adverse effect). Warfarin dosage is an obvious example of this success (Cooper et al. 2008). A GWAS study led to discover a set of SNPs in several genes that influence warfarin dosing. This, with further validation studies, became a clinical genetic test, which allowed physicians to give the correct amount of warfarin to patients.

The relationship between genetic analysis and clinical outcome fostered the field of **personalized medicine**. The current project "10K genomes" in the United Kingdom (Koepfli et al. 2015) is a scientific enterprise taken by the British government for a personalized medicine in the public health care. The objective is to diagnose patients with rare diseases, who otherwise would never get proper treatment. Candidate genes detected through GWAS are generating large datasets of genetic variants associated with disorders, which are being deposited in public databases, such as Online Mendelian Inheritance in Man (OMIM) (Scott et al., n.d.), the database of Genotypes and Phenotypes (dbGAP) (Tryka et al. 2014) or Humsavar (The UniProt Consortium 2007).

## 1.3 Interaction networks and pathways in the cells

The analysis of the data obtained by high-throughput technologies also produced a revolution in the biological field. It marked the start of the “OMIC era” (Kandpal et al. 2009). Genome, Proteome, Peptidome, Exome, Transcriptome, are different ways to profile and classify the biological activities of the cell. However, the analysis of any of these profiles in isolation does not give the answer to fundamental questions about the genotype-phenotype relationship (Vidal, Cusick, and Barabási

2011). To infer the physiological effect caused by the changes in these profiles is necessary to study how the elements of a cell affect each other. In fact, many of such “omic” sciences are inherently a system-based science that requires an integrated approach to study the elements on a given condition by analyzing the interplay between these elements to achieve a biochemical function in the context of a network (Wu, Hasan, and Chen 2014). The signaling pathways of the cell constitute a well-understood example of how the elements of the cell interact to elicit a molecular process. From an outside stimulus, receptor proteins transduce the signal using small molecules known as second messengers, such as the circular Adenosine Monophosphate (cAMP). Enzymes like kinases use the energy stored in Adenosine Triphosphate (ATP) to activate other proteins and start a cascade that produces the release of other second messengers, like Inositol Triphosphate (IP3) and calcium ions. Second messengers can be sensed by other proteins to inhibit the signaling or to start other pathways, in many cases reaching the nucleus and regulating the DNA transcription (Lemmon and Schlessinger 2010). Pathways become interconnected networks when components of one pathway interact and control elements of another pathway. Graph theory can help to analyze a system as complex as the cell. A young discipline in biology, s**ystems biology**, is taking advantage of computational approaches to understand how these interactions can have a response (Ma’ayan 2009). Systems biology is the study of how molecules interact to give rise to subcellular machineries that form the functional units capable of performing the physiological functions needed for the cell, tissue or organ (Bhalla and Iyengar 1999). The network analysis in systems biology intent to gain biological meaning using a global network diagram derived from available data (Wu, Xiaogang, and Chen, n.d.)

Large-scale studies at proteomic level have become widely accessible to the community (Kuhner et al. 2009, Gavin et al. 2006, Yu et al. 2008) and are generating a diverse and increasing amount of data, including protein binding and pathway information (Aranda et al. 2010, Szklarczyk et al. 2015, Ogata et al. 1999). This has facilitated the computational construction of genome-wide networks of interactions, or "**interactomes**" (Rolland et al. 2014)**.** Thus, a system-wide approach can point out the essential elements for regulating a given biological process (Wu, Hasan, and Chen 2014). For example, the response to a stimulus depends on the state of the signaling networks, and this can be used in system biology to predict the outcome of such stimulus at molecular level (Janes et al. 2005). An interactome network describes the interaction of genes or gene products, which means that to provide some explanation of the genotype-phenotype relationships the networks have

to include interactions at different levels. To make the predictions reliable and unbiased, the macromolecular interactions such as DNA-protein, post-translational modification and its target, or protein-protein interactions (PPI) need to be of high quality and extensive (Rolland et al. 2014). PPIs are probably the most critical networks as they underlie in almost all key cellular events like proliferation, cell signalling, regulation or cell morphology alteration (Teichmann 2002).

The most widely-used high-throughput laboratory techniques to construct PPI networks are perhaps the Yeast Two-hybrid (Y2H), and Tandem Affinity Purification coupled with Mass Spectrometry (TAP-MS). Y2H is an ingenious system that uses separable transcriptional factors and a reporter gene to prove the interaction between two proteins. The transcriptional factors have two separable domains, a DNA-binding domain (BD) and a transcription activation domain (AD). The target protein is fused with the BD and is called the bait, the binding partner is fused with the AD and is called the prey. The interaction between bait and prey reconstitute the function as a transcription factor, which can allow the expression of reporter gene downstream from the AD binding sequence (Fields, Stanley, and Ok-kyu 1989). TAP-MS relies on tags attached to the N-terminus of target proteins. The intended target proteins are expressed inside the cell and allowed to interact. Then, the target protein complexes are isolated by two steps of affinity purification. The proteins that co-purified with the tagged proteins are identified by mass spectrometry (Puig et al. 2001). Complementing the initially constructed networks with text mining of the literature has facilitated building the interactomes of different organisms, like S.cerevisiae (Uetz and Hughes 2000; Ito et al. 2000), C. elegants (S. Li et al. 2004), A. thaliana (Cui et al. 2007), D. melanogaster (Giot et al. 2003; Guruharsha et al. 2011) and human (Ewing et al. 2007).

The estimated size of the human interactome ranges from 130,000 to around 650,000 binary protein-protein interactions (Rual et al. 2005; Stumpf et al. 2008). Currently, the high confidence human interactome accounts for ~14,000 PPIs (Rolland et al. 2014), far from being completed. The main challenge in the study of the interaction networks is to extract biologically relevant information from an extensive list of interactions taking into account different sources of the data, in order to gain insight into the molecular mechanism that drives various conditions (Khatri, Sirota, and Butte 2012; Glazko and Emmert-Streib 2009). Comprehensive integrative approaches that take into account data from DNA microarrays, protein expression, PPI information, and interaction with metabolites are added to the complexity in the analysis of cellular functions (Ideker et al. 2001;

MacBeath, Gavin 2002). To gain knowledge from this vast source of information, **network and pathway analysis** can help to interpret the changes in the PPIs caused by external stimuli. The first generation of PPI human interactomes allowed network-based answers to the genotype-phenotype relationship, however, given their limited quality were not useful to make global, accurate interpretations (Stelzl et al. 2005; Rual et al. 2005). **Network analysis** uses the topology of the network to highlight key nodes and strong interactions between different molecules, known as modules (G. Li et al. 2014; Hartwell et al. 1999). In network analysis, biological networks are described as “small world and scale-free” networks (Barabási and Oltvai 2004). This basically means that the human interactome contains several highly connected molecules, i.e. nodes that are known as “hubs.” These proteins usually have a fundamental role in signaling pathways and their function is almost essential for the cell. The highly dynamic character of the interactions in the signaling pathways is a characteristic that provides robustness to the interactome (Albert, Jeong, and Barabasi 2000).

In complex networks like the human interactome, there are no clear clusters because of the scale-free property. The scale-free property makes biological networks similar to nonlinear problems like chaos, phase transitions, and fractals (Strogatz 2001). In fact, using only topological information and a nonlinear dynamical modelling known as the ant colony optimization, revealed fractal-like patterns in protein interaction networks in yeast (Wu, Xiaogang, and Chen 2012), Breast Cancer (Wu, Harrison, and Chen 2009), and Alzheimer disease (Wu et al. 2009).

This indicates the complexity of the PPI networks due the dynamics in the cell change in a continue manner. On the other hand, we know that activity in a cell emerges from functional modules, defined as a group of different proteins that interact but that are not necessarily present in the same space and time (Hartwell et al. 1999; Pizzuti and Rombo 2014). Thus, there must exist some degree of clustering. There are two different ways to detect functional modules: graph clustering, or distant-based clustering. Graph clustering takes full advantage of the topology itself, as it searches for groups of nodes in the network that have more intra-connections than inter-connections. Some graph clustering methods are Highly Connected Subgraph (HCS) (Hartuv, Erez, and Ron 2000), Restricted Neighborhood Search Clustering (RNSC) (King, Przulj, and Jurisica 2004) and Markov Clustering (MCL) (Enright, Van Dongen, and Ouzounis 2002). In the distance-based clustering method, some metrics from graph theory become the similarity measure that

clustering algorithms will use to identify the modules. Some of these metrics are the number of edges (Vazquez et al. 2003), shortest path (Arnau, Mars, and Marín 2005), and shortest path profiles (Maciag et al. 2006).

Parallel to the network analysis, **pathway analysis** is a simplified approach that reduces the complexity of interpreting all available data and increases the explanatory power. Grouping proteins, genes, and PPIs according to the biological process where they participate can reveal clustering for a given event. This categorization breaks down long lists into smaller subsets that can be used to identify differences between two conditions, thus increasing the explanatory power (Khatri, Sirota, and Butte 2012; Glazko and Emmert-Streib 2009). Pathway analysis is different from the network analysis, because it uses functional information about the proteins, like cellular localization, catalytic activity, and processing aspects. Pathway analysis is more successful when it includes PPIs networks, Gene Ontology terms (GO) and expression data. The assumption that proteins in the same pathway and with common functions are tightly regulated can lead to the discovery of the "pathway network module". In this way, we can delimit a large set of proteins that co-regulate each other to perform a particular cellular function (Wu, Hasan, and Chen 2014). Additionally, in some biological networks, there is a correlation between GO terms and node distance (Sevilla et al. 2005; Lord et al. 2003; Cho et al. 2007). On the downside, the annotation of a GO term has a heterogeneous origin, based on a variety of experiments and computational methods, which often leads to inaccurate/contradictory annotations and interpretation problems due the functional diversity of the proteins under different conditions (Luciani and Bazzoni 2012).

There are different databases for protein networks and biological pathways: Biogrid (Chatr-Aryamontri et al. 2015), Reactome (Croft 2010), KEGG (Qiu and Yu-Qing 2013), STRING (Mering 2003), PAGED (H. Huang et al. 2012), HPD (Chowbina et al. 2009), BioCarta (Nishimura and Darryl 2001), or Interactome3D (Mosca, Céol, and Aloy 2013). Many of these databases provide, in addition to the list of interactions, information like the effect of the interaction (inhibition or activation), or the location of the interaction (e.g., nucleus, cytoplasm, and so forth). On the other hand, a number of databases provide experimentally obtained structures of PPIs but lack the integrating context of the networks: 3D interologs (Lo et al. 2010), 3D complex (Levy et al. 2006), SCOPPI (Winter et al. 2006), IBIS (Shoemaker et al. 2012), 3did (Mosca et al. 2014), PIFACE (Cukuroglu et al. 2014). Interestingly, STRING and Interactome3D provide the 3D structures of the

proteins and the complexes they form, in the context of network data.

## 1.4 Defining diseases as PPI networks

Smaller subsets of the human interactome can be used to find answers to the genotype-phenotype relationship. Combining GWAS data, technically a “cause-effect” list for genes, with the network view has provided the most comprehensive data for complex diseases. As complex diseases are caused by several genes (e.g., heart disease, cancer, and diabetes), the use of networks seems a natural approach to gain insight on their molecular bases. **The human diseasome,** which links phenotypic features to all known disease genes, is the result of that approach (Goh et al. 2007). The human diseasome can be exemplified by a bipartite graph in which a set of disease nodes is linked with disease gene nodes (Goh and Choi 2012). The objective of the construction of a network for each complex disease holds the promise of identifying those interactions altered by mutations, which could help to find a treatment to revert the network back to normal state. The core of the human diseasome can be identified using a set of PPIs that are affected by a mutation leading to a pathological state. It can be obtained by purely computational tools and can help to highlight the key players that drive most of the characterized diseases (Janjić, and Nataša 2012). Even if the main disease-related proteins are identified, these advances do not mean a way to find a magic bullet for all pathologies. The highly dynamic nature of the signaling pathways due to their inter-connectivity is a characteristic that adds robustness to the cell (Kitano 2004). One example of a robust disease is cancer. A cancer tumor is a population of different cell types, each harboring their own mutations (Calon et al. 2012; Ding et al. 2012; Gerlinger et al. 2012; Hou et al. 2012). In this way, there are intracellular and intercellular interaction networks with different dynamics, since not all the proteome is expressed sequentially in a specific cue (S. P. Shah et al. 2012). Given the finite number of interactions between nodes in the cellular networks, there is a limit to the number of network configurations or states they can adopt. By rewiring the connections of a signaling network, cancer mutations are probably creating new states that are only present in cancer cells, and that are known as cancer network attractors states (Creixell et al. 2012).

The inter-connectivity of signaling pathways or pathway crosstalk is the underlying reason for such high network dynamics and is one of the reasons why a drug specifically designed for a key protein in a disease can fail. Thus, when a key pathway is inhibited, the cell may use another pathway that can have a similar physiological effect. The multiple layers of gene regulatory

interactions modified by the alteration of the genetic material and structure (e.g. mutations in DNA, or aneuploidy at chromosomal level) combined with feedback loops give rise to the robustness of the cancer cell. Thus, ‘de novo’ mutations during chemotherapy, in combination with feedback controls, allow the cancer cell to be resistant to treatment (Kitano 2004).

This is a problem from a pharmaceutical point of view, since a designed drug will be labeled as useless when it fails to stop the disease progression. Traditionally, the pharmacological approach to treat a disease has been a reductionist one, i.e. “one disease - one target - one drug”. In recent years, this has caused two major problems in the pharmaceutical field: 1) “me-too” drugs, when many companies design drugs for the same targets, and 2) poor assignment of medication to phenotypes due to multi-target properties (Barratt and Frail 2012). The combination of systems biology with drug discovery, known as **network pharmacology,** is starting to change the approach of “one disease - one target - one drug” (Brown and Yasushi 2012). The generation of diseasome networks does not aim exclusively to determine the role of the gene or protein. We can add information such as the mutations that cause a given disease or confer susceptibility to a drug, in order to determine the role of individual players in the crosstalk context. A recent study showed that by using the pathway crosstalk data and available approved drugs it is possible to combine certain drugs targeting a particular signaling pathway in order to reduce the dose, while still being effective against cancer. As a consequence, this strategy has helped to develop an effective treatment less harmful to the patient (Jaeger, and Aloy 2012; Jaeger et al. 2015).

Progress made with these different approaches has improved the rational design of drugs. Most of the designed drugs aim to block the binding sites of a protein. If the expected target of a drug is an enzyme, a first approach is to block the catalytic binding site, as in the case of neuraminidase inhibitors (Russell et al. 2006; Vavricka et al. 2011). An alternative approach to target protein activity is by interfering protein interaction binding sites, therefore stabilizing or disrupting PPIs, like the transthyretin inhibitors (Gallego et al. 2016). In fact, some mutations are lethal by modifying or interfering in a protein binding site, as in the case of the formation of amyloid fibrils that precedes the Amielod Lateral Sclerosis or Alzheimer's disease. In these cases, a mutation in the protein transthyretin destabilizes the formation of the normal multimer protein state, causing the proteins to aggregate in the form of fibrils. In this way, the mechanistic detail of how the protein is affected by drugs or mutations can only be given by the 3D structure of the protein

and the complexes that it forms. Therefore, a high-quality image of the 3D structure of the proteins and the complexes they can form is an essential requirement for the design of effective drugs, which combined with the network approach, gives rise to new pharmacological strategies to treat disease in humans.

## 1.5 Determination of the three-dimensional protein structures

Several diseases such as cancer or RASopaties (a group of diseases related to the malfunction of Ras signaling pathway), display altered PPIs networks (Kiel and Serrano 2014). Current therapies that only target a single protein are not efficient in restoring the phenotype to normal in intricate signaling pathways. It would be needed to use a network-based therapeutic strategy to turn back the appearance of a malignant attractor state in the signaling network (Vidal, Cusick, and Barabási 2011). The use of pathway analysis on the network of interest could help to force the regression to the normal state. Current network maps give information on the relationships of genes or interactions between proteins. However, the vast majority neglects the structural information provided by repositories like the Protein Data Bank (PDB) (Kiel, Beltrao, and Serrano 2008). Databases such as STRING (Mering  2003) and Interactome3D (Mosca, Céol, and Aloy 2013) give information about the reliability of the interactions for a given protein, and if available, they provide the 3D structure of the complexes it can form. This type of information is of paramount importance for the rational design of drugs or repurposing studies. Unfortunately, there is a big problem when it comes to the use of 3D structures for PPIs. While there are 3D structures for nearly 50% of the proteins forming the human proteome (Müller, MacCallum, and Sternberg 2002), only 7% of the complexes forming the known human interactome is structurally characterized (Mosca, Céol, and Aloy 2013). Cheap and massive sequencing technologies have provided drafts of complete genomes, and mass spectroscopy the identification of thousands of proteins. However, obtaining the atomic resolution of a protein is a slow and arduous process. Below are detailed the major experimental and computational approaches to protein complex structures.

### 1.5.1 X-ray Crystallography

The most widely used and accurate approach for obtaining high-resolution protein structures is the crystallography of proteins in combination with X-ray diffraction. A highly concentrated purified protein is needed for crystallization. Exposure of the crystal to an x-ray beam provides a diffraction spot pattern that gives information about “structures factors”, which allows building a map of

electron density. The mathematical process to convert the intensities of the diffraction spots to the electron map is known as the phase resolution problem. The goal is to build a model of the protein based on this map, in which the protein sequence is the input to produce a thermodynamically stable structure (Smyth 2000). However, the process is very slow, requires a large amount of sample at a high purity/quality, and often the protein has to be modified to achieve crystallization, with the risk of modifying the natural folding of the protein. Obtaining a crystal is not a routine process, since the conditions to find the formation of a crystal vary from sample to sample. Even after successfully obtaining a crystal, it might not be sufficiently optimal to determine the structure with high definition. Moreover, factors like the temperature and pH can affect the folding of the protein so that different structures can be obtained (Schiffer et al. 1989). In fact, there are cases where the applicability of this technique is extremely hard or unfeasible. Membrane proteins, low affinity complexes fall in this categorization since obtaining a crystal requires the stabilization by the membrane bilayer or a chemical scafold to maintain the proteins folded and in close contact altering their natural conformation. Also, intrinsically disordered proteins, or very flexible loops present a problem since the periodicity required in for solving the phase problem can not be achieved. Aditionally, the use of crystal as the representation of the *in vivo* conditions or the biological relevant conformation of the protein has been challenged and still under debate. (R. P. Bahadur and Zacharias 2008; Ranjit Prasad Bahadur et al. 2004; Ofran and Rost 2003).

### 1.5.2 Nuclear Magnetic Resonance (NMR)

Another widely used technique to elucidate the 3D structure of a protein is **Nuclear Magnetic Resonance (NMR)**. Since the 50’s NMR has evolved from the field of physics to the medical application. NMR relies on the use of strong magnetic fields where the nuclei and electrons of the atoms absorb the electromagnetic energy and reach a frequency of emission similar to the natural isotopes (typically $C^{13}$ and $H^{1}$). However, this signal changes due the surrounding environment, thus giving also information of the nearby atoms. The advantage of NMR over the crystallography is that protein is in solution, a more natural environment that allows small movement of the proteins. It is very useful for determining the motions of proteins, including those large portions that do not have specific folding and are called intrinsically disordered. NMR experiments are time consuming and expensive, since larger molecules need machines with higher and higher frequency magnets. Thus, a major drawback of NMR is the size of the sample, since currently structures larger than 35 kDa cannot be determined. Therefore, in comparision with X-ray crystallography, very few complete

structures of PPIs have been obtained by NMR, being especially difficult the case of multi complexes (Marion 2013; N. Shah et al. 2006)

### 1.5.3 Cryogenic Electron Microscopy (Cryo-EM)

This technique is based on Electron Microscopy (EM). Standard EM needs to coat the sample with some special protector that usually contains metal particles like silver or gold, generating a layer with valleys and mountains according to the shape of the sample. Then, a laser is applied to the surface produced in the layer, creating the image in slices as it passes like in confocal microscopy. However to enhance the image of minuscule samples, and to prevent degradation, and motion, the sample is fixed on a plate at very low temperatures, which is the basis for Cryo-EM.

Until recently Cryo-EM was regarded as a low-resolution technique because it presented a barrier at 6 Å of resolution and only allowed the inference of huge structures. However, with the recent improvement of the sensors, and high-level algorithms for image recognition, the reconstruction of the 3D structure up to 2 Å resolution is possible. (Dominika, and Hans 2015).

### 1.5.4 Small angle X-ray scattering

A recent structure determination development is the small angle X-ray scattering (SAXS). In contrast to crystallography, in SAXS the sample is exposed to an X-ray beam of a particular wavelength that is moved from 0 to 5 degrees to produce intensity distributions. The generated profile contains structural information of the atoms in the protein that can be in three different regions: the Guinier region that can be related to the average size of the group of atoms, the Fourier regions that contain information about the shape of the atoms in the protein, and the Porod region that provides information about the surface occupied in the volume by the atoms (Boldon, Laliberte, and Liu 2015). The advantage of this method is that proteins can be studied in different media and even disordered. Interestingly, for the resolution of protein complexes, this technique can be coupled with other computational methods such as molecular dynamics or protein docking algorithms (Jiménez-García et al. 2015).

### 1.5.5 Computational modeling

Despite all the recent advances, the majority of protein complexes are yet to be resolved. Thus, an option to fill the structural gap in the human interactome is the use of **computational modeling**. The first attempt would be to construct the 3D structure of a complex from the amino acid sequence

based on the available structure of complexes formed by similar proteins, using ab initio or homology-based modeling techniques similar to those used to model individual proteins. In this sense, the CASP experiment (**Critical Assessment of Techniques for Protein Structure Prediction**) (Kryshtafovych et al. 2013) aims to assess how accurate is the prediction of current modeling programs in blind conditions. One approach is to make fully *ab initio* predictions from the protein sequences, considering the physicochemical properties of the amino acid and the energy terms that drive the folding. An alternative approach is to take advantage of the structures deposited in the PDB, by comparing gene products of different genes but with similar folding, so-called homology modeling. Homology modeling is a powerful tool to determine the 3D structure of proteins and complexes with a high degree of similarity. The most successful programs in CASP are multithreading software able to use structures deposited in the PDB, sequence similarity, and a little *ab initio* modeling. Winning strategies in the last editions of CASP are those of I-Tasser (Y. Zhang 2008) and QUARK (Y. Zhang 2014) which are programs that integrate fragment search in the PDB with the identification of basic folds that can be used as templates, and then fragments can be assembled into models of proteins.

## 1.6 Expanding the protein-protein structural studies through the use of docking tools

As above described, experimental determination of the structure of a PPI is highly challenging. Co-crystallizing two proteins is much more challenging than finding the right conditions for an individual protein; NMR has a size limitation, which leaves out mesoscopic protein ensembles, and Cryo-EM is still in development. As a consequence, all these experimental procedures can be defined as low-throughput. These limitations create a gap between the number of new PPIs that are being discovered with high throughput experiments, and the very few 3D complex structures that are being determined. Computational approaches aim to compensate the difficulties in the determination of PPIs structures. However, predicting the 3D structure of the complex formed by two interacting proteins is a very challenging problem. The issue is similar to the structural prediction in individual structures, in the sense that both cases need a description of the physicochemical forces that regulate the interactions between the amino acid residues. Features such amino acid complementarity, electrostatics, steric clashes, hydrophobic effect, or hydrogen bonding, are concepts shared between both problems.

Unlike the problem of protein folding where the degrees of freedom in which a protein sequence can fold makes the space of search extremely large, in complex structure prediction, proteins are assumed to have 3D structure. This means that the search space is a six degree problem (three translations and three rotations), if we do not consider internal movements (rigid-body search). Computational tools such as **protein-protein docking** try to predict *ab initio* the correct orientation of two proteins that interact. Two major technical aspects can be found in the majority of docking methods: the generation of a large variety of structural models (sampling) and the identification of the correct docking poses with a proper function (scoring) (Huang, Sheng-You 2014). At the core of several docking protocols resides the idea of geometric complementarity in the protein-protein interface. However, in recent years different mechanisms have been proposed for protein-protein association:

- The basic mechanism, called “lock and key”, was directly inspired in complementarity, where the unbound monomers have a matching symmetry that is energetically favorable for the complex formation. This binding mechanism implies that both monomers are rigid, and they fit into one another.
- The "induced fit" mechanism involves conformational changes after binding of both monomers, before achieving the energetically favorable formed complex (Kuser et al. 2008).
- The "conformational selection" mechanism assumes that bound states are naturally samples in the individual proteins and the binding partner selects those conformations that are energetically favorable for binding (Gianni et al. 2014).

Protein-protein docking aims to predict the structure of a protein complex, inspired on the association mechanisms above described. In a real case scenario, the only information available is the 3D structure (or a reasonable model) of the unbound proteins. Current sampling strategies can be classified in: exhaustive global search, local shape feature matching, and randomized search.

Exhaustive global search over the protein aims to sample the entire possible space around a protein using as a probe another protein. In a rigid-body assumption, one needs to account for the translation on three axes, and the rotation on three axes, being a six degree of freedom problem. Exhaustive search can be achieved by using a grid to convert the surface of a protein into a coarse description. Then, Fast Fourier Transform (FFT) calculations (Katchalski-Katzir et al. 1992) can be

used to reduce the computational cost by simplifying the translational and rotational search of the molecules. To completely search the 3D space of both proteins, one of the proteins (by convention the biggest one) is fixed and becomes the static molecule, while the other one moves in the 3D space through the FFT-based algorithm. The grid representation of the molecules allows to distinguish between the inside, the surface, and the outside of each protein. The next step is to obtain a correlation score for all the relative translations between the two grids. This correlation score can be calculated on the molecular shape complementarity of the grids, by taking only into account the overlapping between surfaces as defined in the protein grids. After speeding the correlation calculation by FFT algorithms, a scoring function is applied and, this process repeats for each of the rotations of the mobile protein. This performs an exhaustive search of the 3D space of the interacting molecules. This method is by far the most popular one and has given rise to different programs where the differences are the description of the molecules on the grid. Some of these programs are FTDock (Gabb, Jackson, and Sternberg 1997), ZDOCK (Chen et al. 2003), SDOCK (Zhang and Lai 2011), PIPER (Kozakov et al. 2006), MolFit (Redington 1992). A drawback of this type of approach is that it consider both proteins as rigid bodies, therefore, while it is suitable for an initial docking approach, it does not take into account the flexibility of both proteins. In fact, flexibility is one of the major current challenges for all docking algorithms.

Another approach is the local shape feature matching, with problem still remaining withing six degrees of freedom. In this type of approach the molecular surface of both unbound proteins is calculated, which helps to identify binding regions. A segmentation algorithm is used to identify geometric features, such as convex, concave, and flat zones. Then, the molecular shape is represented by a graph in which each node is a representation for a surface region of the protein. The next step is to identify matching surfaces, which is called geometric hashing. Programs like Patchdock (Schneidman-Duhovny et al. 2005), DOCK (Kuntz et al. 1982), or LZerD (Esquivel-Rodriguez et al. 2014) use this type of sampling to produce tens of thousands poses in a fast manner. One of the particular problems of this approach is that the generated docking poses often include many atomic clashes, so additional steps of steric checking, clustering of solutions to avoid redundancy, or refinement are needed.

The third approach in sampling is random search. In this case, it is important to define several starting points and then drive the sampling towards the optimal positions. Some methods

such as ICM disco (Fernandez-Recio, Totrov, and Abagyan 2003), RossetaDock and HADDOCK (Dominguez, Boelens, and Bonvin 2003) use random search as part of their docking strategy. Some algorithms based on random search are inspired by the swarms observed in the birds or insects. The best example of this algorithm applied to protein-protein docking is the Particle Swarm Optimization (PSO) (Clerc 2006). For exploring the energetic landscape, the best energetic complexes are selected, and they are subsequently used as new seeds, with the process iterating until there are no new seeds. During the funnel-like search, the process only keeps the energetically favorable conformations and drive the docking proteins to the optimal matching pose. This type of algorithms can consider the flexibility of the proteins in the final refinement phase, during the minimization, or through normal mode representation of the search vectors. These algorithms are very successful to find near-native solutions, but computationally expensive. One successful example is the program Swarmdock (Moal and Bates 2010).

### 1.6.1 Scoring of docking poses

Many current protein-protein docking protocols are successful if the interacting proteins undergo only small conformational changes upon binding. Even in these conditions, docking algorithms generate a large number of incorrect docking poses, so the aim is to place the near-native solutions as close to the top as possible within a ranked list.. An important part of the success depends on the accuracy of the scoring function used to evaluate the docked conformations, which in turns depends on its capabilities to overcome the inaccuracies of the interacting surfaces and singling out near-native conformations (Halperin et al. 2002; Vajda and Kozakov 2009). Generally speaking, scoring aims to identify the lowest-energy state among the different possible states of a given interaction, and thus, in the case of docking, it should be ideally able to describe the energetic aspects of protein-protein association (Moal and Fernández-Recio 2012). For practical predictions, the energy description of a system is estimated by approximate functions, and a large variety of scoring functions have been used, defined at different resolution levels (atomic or residue) (Tobi and Bahar 2006). Docking algorithms often rely on the geometric complementarity of protein-protein interfaces. The essential zones for binding are often preformed in the interacting proteins (Levi 2010), and as a consequence the interface of a protein complex could be considered an inherent geometric feature of the protein structures. This has made shape complementarity a popular ranking criterion to identify near-native solutions. Still, many protein-protein interfaces are flat, so complementarity alone is not enough to describe the right association mode. This is one of the

reasons why a sampling step based only on geometry criteria often fails to produce correct models. Indeed, the physicochemical nature of the residues has a major role in protein association. Important elements include the electrostatic forces with complementary charges helping to provide the micro environment needed for the interface formation and the correct orientation of the proteins, and the hydrophobic effect with the burial of hydrophobic patches favoring the desolvation of the interacting surfaces (Camacho and Vajda 2001, Camacho et al. 1999). Other factors are van der Waals attraction and repulsion, and hydrogen bonding. However, scoring functions that use energy-based terms to model these effects are not yet accurate enough to reliably select near-native solutions from a pool of decoys, and thus further investigation is required to improve the quality of docking predictions.

Usually sampling and scoring are intimately coupled in a docking procedure. However, in many procedures, scoring is performed independently as a post-docking analysis. Basically, this approach consist in using a scoring function to re-rank the poses generated by a given docking program. This strategy could be considered as a type of refinement of the docking results, but using more sophisticated scoring functions than those used during the search phase. The idea behind post-docking approaches comes from the reasonable success of sampling algorithms to produce at least one near-native solution, also called a hit. In many cases, the in-built scoring function during the docking phase cannot be sensitive enough to place the near-native solution within the top of a ranked list of possible conformations. The computational problem is simplified by detaching the scoring functions from the sampling process, which also adds the possibility of combining different scoring functions. Some examples of post-docking methods are pyDock (Cheng, Blundell, and Fernandez-Recio 2007), ZRANK (Pierce, Brian, and Zhiping 2007), SIPPER (Pons et al. 2011), DARS (Chuang et al. 2008). Given that docking programs typically report decoys ranked with only one or two scoring functions, it remains to be seen whether a given method could further benefit from the accumulated knowledge derived from the variety of currently available scoring functions that have been reported in the literature, many of which were developed for different modeling problems (Tobi 2010). One example of this is the combination PIE/PIER (Viswanath, Ravikant, and Elber 2012). In some methods, the scoring functions are also combined with the inclusion of protein flexibility, like in Fiberdock (Mashiach, Nussinov, and Wolfson 2010), Firedock (Andrusier, Nussinov, and Wolfson 2007), or RDOCK (Li, Rong, and Zhiping 2003).

Among the different scoring functions applied as post-docking analysis, we note the program pyDock (Cheng, Blundell, and Fernandez-Recio 2007), which is an outstanding and consistent protein-protein protocol using the FTDock or ZDOCK sampling combined with a scoring function. The pyDock scoring function is formed by three energy-based terms: Coulombic electrostatics, desolvation energy and van der Waals potential. A protein is a charged entity and its surface has to be in constant contact with solvent molecules, so considering the electrostatic charges of the proteins is the basis of the majority of the scoring functions. But electrostatics alone is not enough to place the two interacting proteins in the optimal position, so there is a need for additional terms to help to improve the algorithm. Since many of the binding surfaces are flat, and the critical contact residues at the interface are often hydrophobic, desolvation plays a major role in creating the micro-environment necessary to allow the formation of a strong interaction between proteins. On the other side, the van der Waals energy is usually important for the final assembly of two given proteins, and it is very dependent on the correct side-chain conformations. When docking is rigid-body, this potential is very noisy. The use of all the above mentioned energy descriptors makes pyDock a very versatile, non-deterministic, and adaptable docking method as I will show in this thesis.

Analogous to CASP, protein-protein docking programs are blindly assessed in the **Critical Assessment of PRedicted Interactions (CAPRI)** (Janin et al. 2003), which is an international scientific effort to boost the development of different approaches to solve the problem of protein-protein docking. After more than fifteen years since the first edition, the CAPRI experiment is now the source of standard protein-protein docking sets and quality measurements.

### 1.6.2 Template-based docking

In addition to ab initio docking, the interface between two interacting proteins could be modeled using the existing structural data in the PDB (Sinha, Kundrotas, and Vakser 2012). **Figure 2** provides a comparison with ab initio docking. As seen in modeling of individual proteins, some evolutionary distant PPIs converged in a structural conformation which is optimal for the recognition. This type of PPIs receives the name of interologs (Matthews 2001). The identification of interlogs facilitates the study of PPIs. The conservation of the structural conformation of the interface through evolution has also demonstrated a plasticity to changes, where only a 66% of the interface patch is conserved leaving a 34% of the patch tolerable to residue changes (Faure,

Andreani, and Guerois 2012). However, the interface is also a dynamic part of the protein that can change during binding (Hamp, Tobias, and Burkhard 2012). In fact, the inclusion of evolutionary data in the context of interface predictions seem to give additional confidence in the prediction (Hamp, Tobias, and Burkhard 2015; Katsonis and Lichtarge 2014).

It has recently claimed that there could be available templates for most of the known protein complexes (Kundrotas et al. 2012). However, in the case of remote homology, i.e. the twilight zone, the available templates do not provide better modeling than *ab initio* docking (Negroni, Mosca, and Aloy 2014).

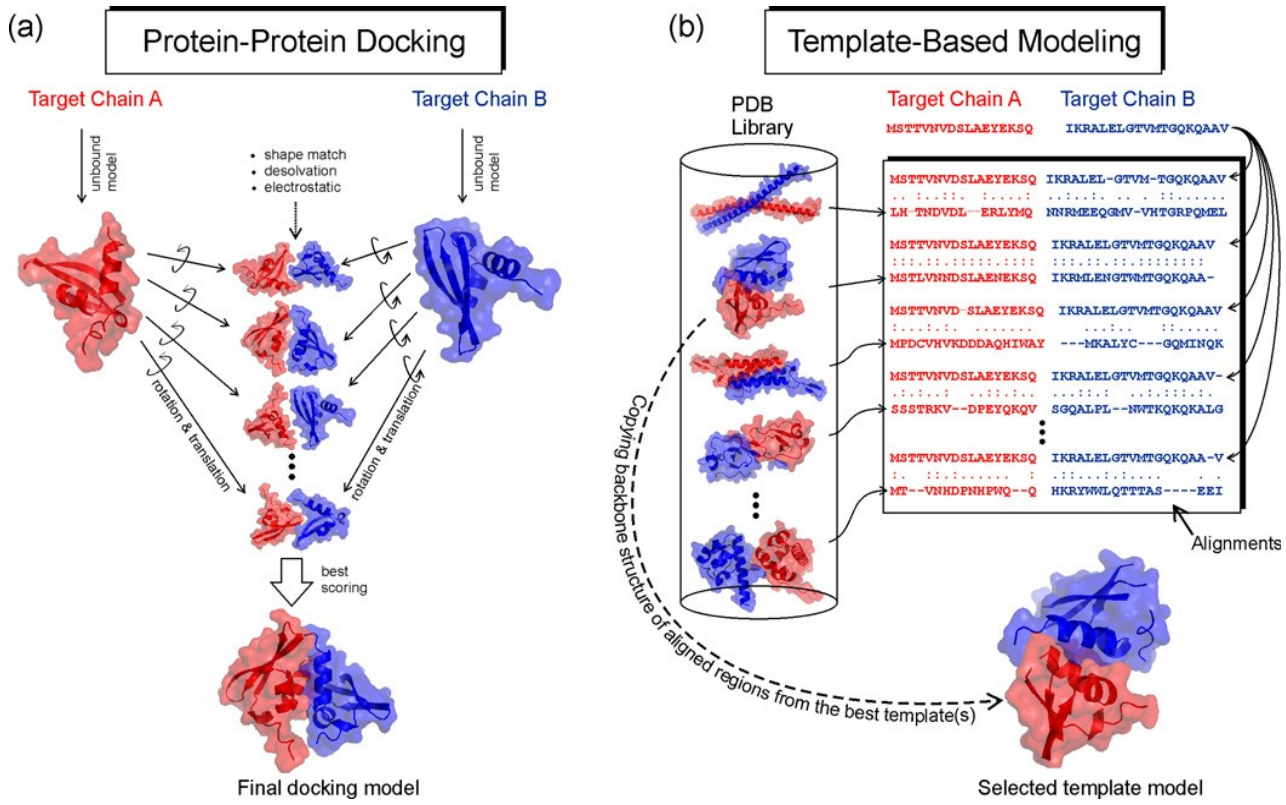



*Figure 2:* ***Comparison between protein-protein docking and template-based docking (from Szilagyi and Zhang 2014)***

## 1.7 Interface and hot-spot prediction

The use of new approaches continues to enable the study of protein interactions from different perspectives. The protein-protein interface is a critical zone for molecular recognition, formed by an average of ~28 residues, accounting for around 1000 $A^2$ of the area in one protein, and mostly

flat(Levy 2010). Based on the relative Accessible Surface Area (rASA) of the residues in the interface, three different zones could be defined (Levy 2010):

1) The core, formed by residues that are exposed in the monomers and form the necessary contacts for the interaction, contributing the most to the binding energy.
2) The rim, which shields the core from the solvent providing the micro-environment required for establishing the interaction.
3) The support, formed by residues that shift from buried to exposed when the complex is formed, helping to establish the interaction.

Usually, in the protein-protein interfaces, there are only a few residues that contribute the most to the binding energy, and are called interaction hot-spots. Alanine scanning (Morrison and Weiss 2001) can be used to experimentally describe the contribution of the different residues to the interaction. The technique consists in performing point mutations in the protein sequence for alanine, so that the chemical neutral nature and size of the alanine allows to mimic the removal of a given residue without perturbing too much the secondary structure. Experimental analyses of hot-spots have found that the core residues tend to contribute the most to the binding energy.

Experimental alanine scanning is time consuming and costly. Computational approaches aim to complement experimental data. Several interface prediction methods have been reported, based on sequence conservation (Valencia and Florencio 2005; Minhas et al. 2013), or on physicochemical properties of the residues (Dong et al. 2007; Grosdidier and Fernández-Recio 2008; Hwang, Vreven, and Weng 2014). This type of methods can be applied for its simplicity in a high-throughput manner to analyze complete interactomes and obtain fast, relevant information for a full set of PPIs.

Interface residues seem to play different roles in disease according to the region they belong. In a recent study found that the core interface residues are more susceptible to disease-related mutations, in contrast to those in the rim regions, which is consistent with the existence of hot-spot residues at the interface of PPIs (David and Sternberg 2015). Complementary work showed that about 11% of all known disease-associated SNPs also land outside but near to the interface (Gao et al. 2015). Both studies found that the residues that are more vulnerable to disease-related mutations

are residues buried in the interface; although they seem to differ about the preferred localization of these mutations.

## 1.8 Protein-protein benchmark sets

In order to assess the predictive accuracy of a newly developed method, it would be necessary to have a reference set widely accepted by the community. In the case of the protein-protein docking field, the reference set needs to have the crystal structure of the proteins in a free state and that of the complexed or bound state. These structures must have a high resolution, and good coverage of the proteins. In addition, the protein set should to be diverse enough, so that it can represent as many as the known protein families as possible. The current version of the most widely used protein-protein docking benchmark has 231 protein complexes (Vreven et al. 2015). Each of those complexes has the crystal structure of the proteins in unbound form and the bound form. The protein docking benchmark is divided into subcategories according to the difficulty, based on the conformational changes that the proteins undergo from unbound to bound states. The most difficult category corresponds to the cases that are the most difficult to predict with current protein docking algorithms, mostly due to the large conformational changes of the proteins.

Other benchmarks have been reported to assess different methods for PPI modeling, like binding affinity changes upon mutation (Moal and Fernández-Recio 2012), scorer sets from CAPRI (Lensink and Wodak 2014), or binding affinity data sets (Kastritis et al. 2011). There are other useful databases such as template libraries (DOCKGROUND; Liu, Gao, and Vakser 2008), structural datasets with similarity between sequences (3D-Complex; Levy et al. 2006), or classification of the domain-domain interaction on protein complexes like SCOPPI (Winter et al. 2006).

## 1.9 Extracting meaningful information from the biological Big Data

As stated in the previous sections, the dramatic drop in costs of genome sequencing is generating an unprecedented amount of genetic data on biological and pathological situations, which together with the complexity of the genotype-to-phenotype information conversion derived from concepts that were discussed in previous sections, such as interaction networks, pathway crosstalk, etc., makes that the idea of “Big Data” is more and more linked with the biological sciences. In these moments, new genome sequencing technologies provide large datasets, which are especially relevant to the

field of healthcare, such as The Encyclopedia of DNA Elements (ENCODE) (Project, Leja, and Birney 2012), or The Cancer Genome Atlas (TCGA) (Project, Leja, and Birney 2012; Zenklusen 2014). The problem of analyzing Big Data in biology has fostered many different international programs, which aim to order, annotate and compare the available data to gain knowledge from it (Marx 2013). Now in biological sciences, there is a large and increasing number of observations generated on different phenomena, which needs the implementation of algorithms borrowed from computational sciences. On the other hand, the study of biological problems often depends on a significant number of variables at different levels of resolution and typology, like the genetic background, temperature, humidity, cell type, metabolite concentration, transcription rates and so on. Nonetheless, the biological data contains an enormous amount of repetitive patterns, from the regulatory elements that precede a gene, to the folding topologies of proteins, which can help to perform statistical analysis.

Traditional statistical approaches could not be sufficiently accurate when analyzing the data that is characterized by a big number of variables. To deal with this type of problem, computer algorithms are training on a sufficiently large number of examples of these different patterns that are found in the experimental observations, as well as on artificial data that are created to mimic the observed patterns (Hughey, Hastie, and Butte 2016).

Borrowed from the computer science field, data mining techniques have been used in a variety of scientific and social areas, ranging from economy to biology. The reason for a biologist to use data mining schemes is the power of prediction that can be gained by using large datasets as well as the capabilities to identify the relevant features. Data mining and statistics go hand in hand with the need of a well-characterized source of examples for the extraction of the essential features. Data mining applications, usually based on machine learning protocols, can accomplish the analysis of multivariate data and are very efficient in classifying data with many patterns. In most of the computational protocols and techniques it is mandatory to obtain fast, reproducible results. The quality of results are highly dependent on characteristics of the dataset like which features to consider, number of missing observations, or balance between the different classes of examples. With a properly classified and ordered dataset, the resulted trained model can easily find information and relationships that are not evident in a simple initial analysis. Data mining application rely on four different styles of learning:

1) Classification learning. This style depends on a properly detailed set of examples from which it is expected to learn the patterns that will help to classify unseen examples.

2) Association learning. In this case, the learning scheme seeks for any association among features, i.e. not just those that can help to classify the unseen examples, which involves dealing with the problem of summarizing a given set of data.

3) Clustering. Here the scheme seeks for groups of examples that can be classified together.

4) Numeric prediction. In this style the prediction is not a class of the problem but rather a number.

Many biological problems are classification problems, and this is the reason why the classification learning is the most used style in the bioinformatics field (Wang 2014). The classification learning is called supervised because it functions based on the class examples provided, in other words, the way it works is using the examples and the given outcomes to create a model for the predictions, which means that it depends on the size and reliability of the observations used in order to produce a high-quality prediction (**Figure 3**). In some data sets no class values are given, in which case the classification learning is called unsupervised. This category deals with finding new patterns assuming the existence of an underlying structure of the dataset. Since there are no class labels, this category strongly depends on on clustering techniques. In both cases, the selection of features in one of the most important steps in machine learning, given that it helps to build the classifier, or the rules to find the new patterns or information. Therefore, in this kind of approaches the feature selection step becomes crucial, since there is a risk to produce over complicated classifiers that will only work for the set used to train the algorithm.

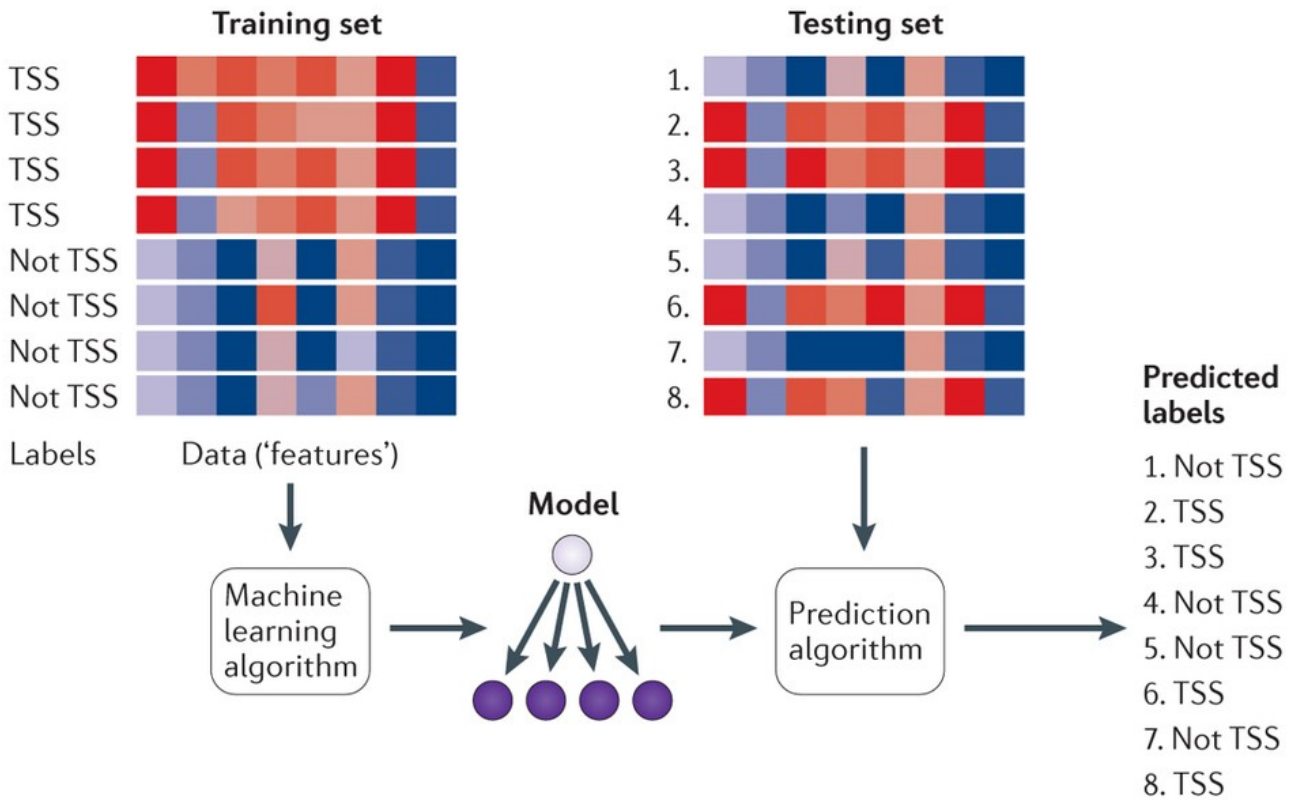



*Figure 3:* ***A general machine learning scheme****.*

The starting point of almost all machine learning protocols is establish a training set to use as examples for the algorithms. After training the researcher can built several modes according to the quality of the data extracted. An external set of unsorted data or test set is used to evaluate the predicted performance of the models. After the test set the result is usually the classification of the test set in an ranked list.

### 1.9.1 Examples of different classifier types

The simplest classifier for numeric features is the linear regression. Any kind of regressions can be applied to classification. This approach is better suited to situations when the class types are binary, e.g. correct and incorrect, 0 and 1, belonging or not, etc. These classifiers search for a weighting factor that will fit the values of the features using a linear function. This function will approximate the "membership" of each feature to the class with a given probability. However, by using a linear function the classifier will assume that the errors are independent and normally distributed with the same standard deviation, and therefore the sum of the estimated probability will become unreal (over 1). This can be avoided by using a logistic transformation that will adjust the probabilities for each target value. Then, weighting factors must be found that fit well the training data.

Most of the research in the biological field deals with multiclass problems, so other machine learning algorithms that can be applied to multiple classes are needed. A popular method to address multiclass problems is the Support Vector Machine (SVM) (Cortes and Vapnik 1995; Joachims et al. 2009). In this approach, the classification is made with a hyperplane that optimally separates the data. The weights assigned to each feature represent an orthogonal vector forming the hyperplane that separates the different classes. One advantage of the SVM algorithm is that the absolute size of the orthogonal vectors gives an indication of how important each feature was for the separation. Another popular algorithm is random forests (RF) (Breiman 2001), which builds a randomized decision tree using a sampling methodology that generates new training sets from the original. This process is called bagging, in which different features of the initial training set are randomly deleted or replicated to create new sets with the same sizes. Bagging generates a diverse ensemble of classifiers by introducing randomness into the learning algorithm's input, often with excellent results.

### 1.9.2 Feature selection

Trying to optimize all the weights for the full set of features can be computationally expensive. As previously discussed, one of the objectives of the feature selection (FS) is to avoid the overfitting of the classifier, especially in supervised learning. There are other benefits of FS. For example, the optimal parameters for a subset of features might not be the same than those for the full feature set (Daelemans et al. 2003). Then, by studying the most important features, we gain understanding on the underlying process that generated that data. FS also lowers the computational cost by selecting the most important features. Note that not all machine learning algorithms have an embedded way to do FS. For instance, filter techniques or wrapper methods can be applied to evaluate features and present a selection as input to build the classifier. Filter techniques assess the relevance of the features according to the properties of the data, but it does not take into account the dependencies between features. Ignoring these dependencies may lead to bad classification performance. In the case of wrapper methods, a different subset of features is used to generate and evaluate different models. In this way the interaction between features is included in the FS, but this increases the possibility of having an overfitted model.

### 1.9.3 Applications of machine learning in biological sciences

The machine learning protocols have already been applied to a large variety of problems in the biological fields. One of the first applications was to predict the probability of different mutations found in a sequenced genome to cause a disease phenotype (Adzhubei et al. 2010). The protocol consisted in training a SVM algorithm, with various features like B-factors, changes in the accessible surface area, and changes in the residue side-chain volume, among others. In this way, the program could identify potentially harmful SNPs in a gene sequence. Other early example is the prediction of a gene splice site (Degroeve et al. 2002), where feature relevance was measured with the use of an SVM, helping to discriminate this structural element recognition from the sequence. In high-throughput experimental data like microarrays, the most informative subset of genes can be isolated using RF algorithms (Moorthy and Mohamad 2011; Díaz-Uriarte and Alvarez de Andrés 2006).

In systems biology, the topology generated by the interaction elements can also become source material for predictions from computer trained programs. Since protein connections change over time or under different conditions, it is possible to couple expression information, like the one provided by RNA-seq experiments, and a Boolean genetic model in order to search for highly connected elements (Crespo et al. 2013).

Trained classifiers and computer trained programs from different types of sources are regularly used in modeling protein-ligand and protein-small molecule interactions, especially for finding the more useful features to detect a suitable binding drug (Myint et al. 2012).

Machine learning protocols can be easily applied to protein-protein docking. Using the standard benchmarks and quality measures widely accepted by the docking community, like those in CAPRI, the programs generate thousands of different poses that include only a handful correct poses. Using a good ranking of these poses generates enough input to train an SVM model to find binding sites for PPIs (Hwang, Vreven, and Weng 2014; Minhas et al. 2013). Another method that combines machine learning and protein docking has been applied to build a predictor for damaging nsSNPs that disrupt the interface of a complex (Goodacre et al. 2016).

“A wise man can learn more from a foolish question
than a fool can learn from a wise answer.”
– Bruce Lee

# Chapter 2 Objectives

The general goal of this thesis is the application of docking scoring methodologies to the characterization of missense mutations in proteins that could be relevant for understanding pathologies in a network-based context. The first part is focused on the analysis of a large number of scoring functions and their performance on different docking methods. In the second part, docking-based methodologies are applied to predict the involvement of missense mutations in protein-protein interactions, and how this can be used to characterize such mutations when there is neither structural information nor knowledge of which complexes are affected.

The specific objectives of this thesis are:

1) Analysis and optimization of current scoring functions for protein docking, and the effect of protein flexibility and binding affinity on the predictions.

2) Optimization of the scoring of docking poses for the identification of interface and hot-spot residues, and its application to characterize missense mutations.

3) Characterization of missense mutations by interface and hot-spot predictions in selected disease-related interaction networks for which little structural data on the interactions is available.

4) Large-scale annotation of missense mutations and their involvement in protein-protein interactions based on computational docking and scoring calculations .

“The harder you work, the harder it is to surrender.”

- Vince Lombardi

# Chapter 3 Methods

## 3.1 Protein-protein docking

In the present work, we have used several well-known and freely available rigid-body protein-protein docking programs, which we ran with the specifications described below (default parameters otherwise). We ran FTDock 2.0 (Gabb, Jackson, and Sternberg 1997) with electrostatics on, grid cell size of 0.7 Å, and surface thickness of 1.3 Å, with a total of 10,000 docking poses generated for each case. We reconstructed of missing side chains of interacting proteins with scwrl 3.0 (Canutescu, Shelenkov, and Dunbrack 2003). ZDOCK 3.0.1 (Pierce, Hourai, and Weng 2011) was used to generate 54,000 docking poses, from which we kept only the highest-scoring 10,000 ones for further analysis. We ran SDOCK (Zhang and Lai 2011) as recommended by their authors, and we conserved 1,000 clustered docking models for further analysis.

## 3.2 Protein-protein docking benchmark sets

We have computed the predictive success rates of the different scoring functions and their combinations on the freely available protein-protein docking benchmark from Weng's laboratory**,** for which the structures of the unbound monomers and the bound complex are available. The docking benchmark version 4.0 (BM 4.0) contains a total of 176 targets (Hwang et al. 2010) while the docking benchmark update version 5.0 (BM 5.0) includes 55 additional targets (Vreven et al. 2015).

For additional validation, we also used a recently published scorers-set benchmark. This benchmark contains 15 published targets from 23 CAPRI assessments. The scorers-set benchmark (Lensink and Wodak 2014) contains more than 19,000 protein complex models generated by 43 different predictors groups, including web servers. Only 10% of them are docking models of acceptable quality or better with a range of 281 to 2182 decoys per case; being the number of acceptable quality decoys 835, medium quality decoys 784, and high-quality decoys 479.

## 3.3 Evaluation of docking predictions

In order to evaluate the predictive success rate of each docking method on the BM 4.0 and BM 5.0, CAPRI quality measurements were calculated for each of the generated structures, based on the fraction of native contacts ($f_{nat}$), interface RMSD (IRMSD) and ligand RMSD (LRMSD) as defined by CAPRI (Lensink, Méndez, and Wodak 2007) with respect to the known reference complex structures. According to CAPRI criteria, the quality of the structures are classified as follows: incorrect [ $f_{nat} < 0.1$ or (LRMSD > 10 Å and IRMSD > 4 Å) ], acceptable [ [$(0.1 \leq f_{nat} < 0.3)$ and (LRMSD ≤ 10 Å or IRMSD ≤ 4 Å)] or [$(f_{nat} \geq 0.3)$ and (LRMSD > 5 Å or IRMSD > 2 Å)] ], medium [ [$(0.3 \leq f_{nat} < 0.5)$ and ( LRMSD ≤ 5.0 Å or IRMSD ≤ 2 Å)] or [$f_{nat} \geq 0.5$ and (LRMSD > 1.0 Å and IRMSD > 1.0 Å)] ], or high accuracy [ $f_{nat} \geq 0.5$ and (LRMSD ≤ 1 Å or IRMSD ≤ 1 Å) ]. This classification was already provided by CAPRI organizers for the models in the CAPRI score set benchmark. Success rates are defined as the percentage of cases in which an acceptable solution (following CAPRI criteria) is found with the top *N* docking models as ranked by a given scoring function.

## 3.4 Protein-protein scoring functions

We selected 73 scoring functions from the CCharPPI server (Iain H. Moal, Jiménez-García, and Fernández-Recio 2015), as shown in **Supplementary Table 1.** These functions were already described in a previous study (I. Moal et al. 2013). We did not use all the scoring functions provided in the CCharPPI server due to technical limitations of the computing platform employed to perform the calculations. For clarity purposes, the majority of contact and distance-dependent residue-level potentials were originally prefixed with ‘CP_’, while atomic and quasi-atomic potentials were prefixed with ‘AP_’.

## 3.5 Cardinality analysis and combination of the normalized values for re-ranking

For all scoring functions we calculated the set of complexes for which an acceptable or better solution appears in the top 10 decoys when ranked by that function. Then, for each pair of scoring functions (A, B), \we calculated the size of their union (eq. 1) and symmetric difference (eq. 2) sets:

$$|A \cup B| = |A| + |B| - |A \cap B| \quad \text{eq(1)}$$

$$|A \Delta B| = |(A \setminus B) \cup (B \setminus A)| \quad \text{eq(2)}$$

These measures, which combine two scoring functions, give an indication of the extent to which the scoring functions are successful on different subsets of the complexes. We also explored a strategy in which scoring functions are combined not just on the basis of their ability to find top 10 solutions in different subsets of the complexes, but also on different subsets of the decoys as deliniated by the docking algorithm that was used to generate them. To do this, we combine three pairs of scoring functions, where each pair was evaluated and selected on basis of its performance on the decoys generated by each of the three docking methods. We calculated the union cardinalities for the unified pair of scoring functions between the three docking methods, forming triplets of scoring function containing one unified pair from FTDock, one unified pair from ZDOCK, and one unified pair from SDOCK, this way combining up to six different scoring functions together. Here A represents a unified pair of scoring functions that performs well in FTDock, B a unified pair of scoring functions that performs well in ZDOCK and C a unified pair of scoring functions that performs well in SDOCK

$$A \cup B \cup C = |A| + |B| + |C| - |A \cap B| - |A \cap C| - |B \cap C| + |A \cap B \cap C| \quad \text{eq(3)}$$

where A represents a unified pair of scoring functions that performs well in FTDock, B a unified pair of scoring functions that performs well in ZDOCK, and C a unified pair of scoring functions that performs well in SDOCK.

To calculate the success rates of these combined functions, we proceeded as follows. First, we combined different energy terms from pairs or triplets of scoring functions, selected using the above measures. To do this, we first normalized each value using the z-score method (eq. 4):

$$Zi = \frac{x - \mu}{\sigma} \quad \text{eq(4)}$$

where x is the value, μ the average and σ is the standard deviation. The normalized values of the scoring function pairs for a given pose are directly added and used to re-rank the list of the poses generated by each method. For combining triplets of scoring functions, we similarly sum the three

z-scores. Note that this is a naive ranking and no weight optimization was undertaken.

## 3.6 Decoy sets for machine learning

We calculated the scoring functions again in an identical manner to the CCharPPI server. We imputed the corresponding missing values as the mean within the docking method. We also included cluster sizes as descriptors, calculated with the g_cluster tool in GROMACS (Pronk et al. 2013),using single-linkage clustering of ligand C**α** positions after superposition on the receptor, with cutoffs at 0.5Å intervals in the 3 to 7Å range. We used a final set of 91 features in total, during the process we normalized the values of these features as z-scores.

For comparative reasons, we considered only the top 500 structures from pyDock, SDOCK, and ZDOCK. pyDock is our docking protocol that uses the FTDock decoys with comparable performance to ZDOCK and SDOCK. Poses were classified as incorrect, acceptable, medium quality, or high quality, using the CAPRI criteria mentioned before. The following table resumes the number of complexes at least a near-native solution, and a complete descriptor set calculated, for both benchmark sets.

*Table 1: Number of hits produced by each of the FFT-based docking method*

| Method | BM 4.0 | BM 5.0 |
|---|---|---|
| pyDock | 103 | 33 |
| SDOCK | 109 | 32 |
| ZDOCK | 114 | 25 |

## 3.7 Model training, selection, and validation

We used the BM 4.0 for training the machine learning program, and it was randomly split into a training set and selection set with a 2:1 ratio. We trained an ensemble support vector machines (Joachims 2006) (R-SVM), with linear kernel function, and the c metaparameter sampled logarithmically 50 times in the $10^{-4}$ to $10^{3}$ range. The R-SVMs were trained to minimize the fraction of swapped pairs regarding a perfect ranking (see **Figure. 4**), which is analogous to the area under ROC curve for binominal classification, using the n-slack algorithm with shrinking heuristics described by Joachims (Joachims 2005), through the SVMrank program (Joachims 2006). We

repeated the process until we generated an ensemble of 200 R-SVMs for each c value. We applied all R-SVM to their respective selection set and scored (see below). In total, the top scoring n models, up to n=50, were selected for combination using the Schulze method (see below) when applied to the external test set. The results in **figure 12 and 13 and Supplementary Table 2** correspond to c and n metaparameter values found by leave-one-out cross-validation. The method is both insensitive to small changes and robust across a wide range of c and n values.

We used the BM 5.0 as the external set to evaluate the performance of our method. This set was no included in the train set and only evaluated with the Schulze method. When using the BM 4.0 as the validation set, the whole data set is split into training, selection and test set in a 2:1:1 ratio. For each complex in the BM 4.0, the consensus Schulze ranking is applied only using R-SVMs for which that complex does not appear in either the training or the model selection set.

## 3.8 Scoring R-SVM models

The total score S evaluated each R-SVM model, the sum of individual scores for each of the $n_t$ complexes in the selection set, $\overline{si}$, compared to the mean score for that complex across the R-SVM ensemble, $\overline{si}$:

$$\sum_{j=1}^{n_t} si - \overline{si} \qquad \text{eq(5)}$$

By taking the score relative to the mean, the total score reduces biases in the selection set by preferentially favoring R-SVM models which perform well on difficult complexes, those which the other models struggle to perform well on, and disfavors models which perform poorly on easy complex, those which the other models do perform well on.

For calculating the individual scores of a complex, the decoys are first clustered at 3.5Å (see above). The rank of the top best decoy list the nc clusters. The overall rank for the complex, r, is the rank of the top listed cluster for which the top-ranked decoy have at least an acceptable quality. Calculation of the scores is as follows:

$$\overline{si} = \frac{\log_{10}(n_c) - \log_{10}(r)}{\log_{10}(n_c)} \qquad \text{eq(6)}$$

This score can range from 0, if only the last cluster has a top ranked decoy that is not

incorrect, to 1, where the top-ranked decoy of the top ranked cluster is at least acceptable. If no acceptable or better solutions appear as top-ranked decoys within any cluster, $\overline{si}$ is set to zero. The logarithmic form gives greater importance to higher ranks heighten the difference in $\overline{si}$ between ranks.

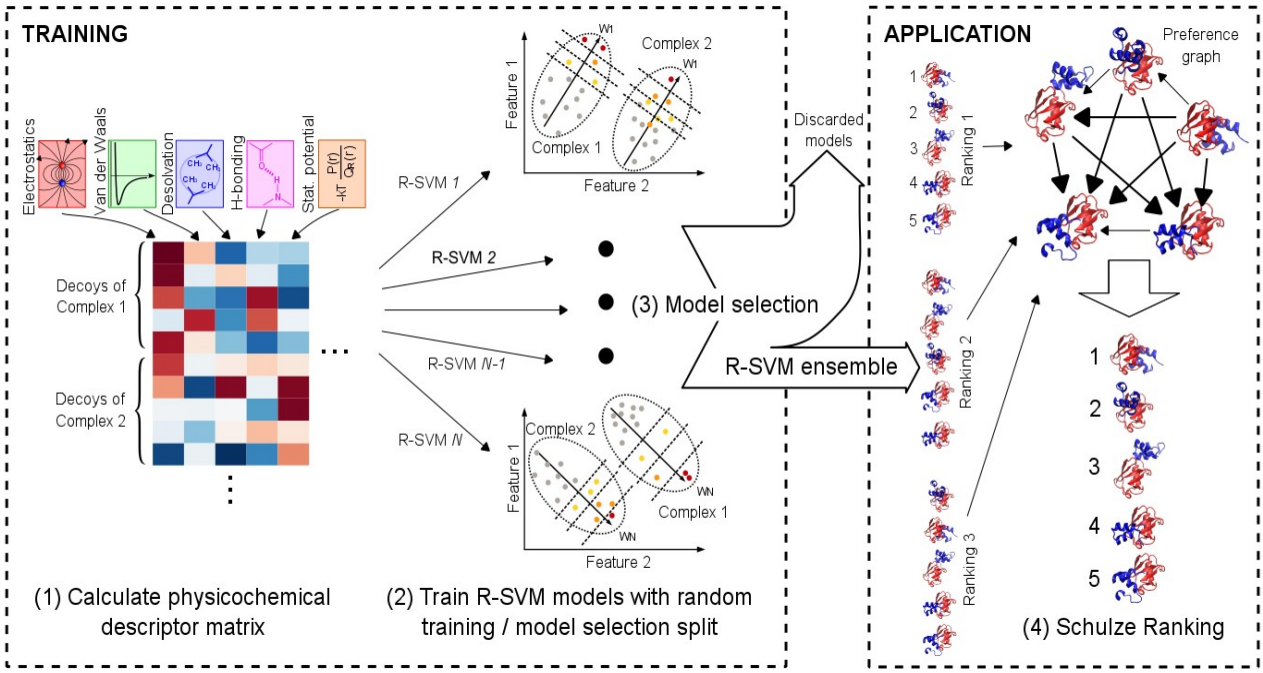



*Figure 4:* ***Scheme of the machine learning protocol and democratic ranking***.

An overview of the algorithm. 1: The training decoys are characterized using physicochemical descriptors, which are organized into a matrix. 2: R-SVMs are calculated with random 2:1 model training and model selection split. Each R-SVM generates a weight vector (w) in descriptor subspace such that decoys for each complex (ellipses), when projected upon it, are ordered to minimize the number of swapped pairs relative to a perfect ranking: high quality (red) > medium quality (orange) > acceptable (yellow) > incorrect (gray). 3: The highest performing models are selected according to their model selection score, and form the R-SVM ensemble. 4: When applied to a new set of decoys, rankings from each R-SVM in the ensemble are combined into a preference graph, with arc (edge) weights indicating the number of times each pose (node) is ranked higher than each other pose. For each pair of poses, a pairwise ranking is obtained by finding the strongest directed path between them, from which the final consensus ranking follows

## 3.9 Applying the method with Schulze ranking

We calculated the physicochemical features for each decoy, to implement the model to the external test sets or new docking cases. The decoys are ranked using each of the n selected models in the R-

SVM ensemble, by their order when projected onto the R-SVM weight vector line in descriptor subspace. Each of these rankings is combined using the Schulze electoral voting system(Schulze 2011).

Complete digraphs, where nodes are decoys and arc weights indicate the number of times the tail node is ranked higher than the head node, are used to find strongest paths between all ordered pairs of decoys, (a,b), where path strength corresponds to the minimum arc weight in a directed path originating at a and terminating at b. Decoy a is ranked higher than decoy b if the strongest path of (a,b) is higher than that of (b,a). As preferences are transitive, a consensus ranking follows directly from the pairwise rankings.

## 3.10 Prediction of extended interface patches by pyDockNIP.

We have developed a new version of the pyDockNIP method for predicting hot-spot residues in a given protein-protein complex, as follows. Docking simulations were run with FTDock (Jackson, Gabb, and Sternberg 1998) to generate 10,000 rigid-body docking poses, which were rescored by pyDock scoring function (Cheng, Blundell, and Fernandez-Recio 2007). From the docking results, normalized interface propensity (NIP) values per residue were calculated with the built-in pyDockNIP module (Grosdidier and Fernández-Recio 2008), and those residues with NIP value greater or equal to 0.2 were predicted as interface hot-spot residues. The novelty here is that the predicted interface patches were extended by including surface residues (relative accessible surface area rASA > 1 $Å^2$) within 10 Å distance from the predicted interface hot-spot residues (**Figure 5**).

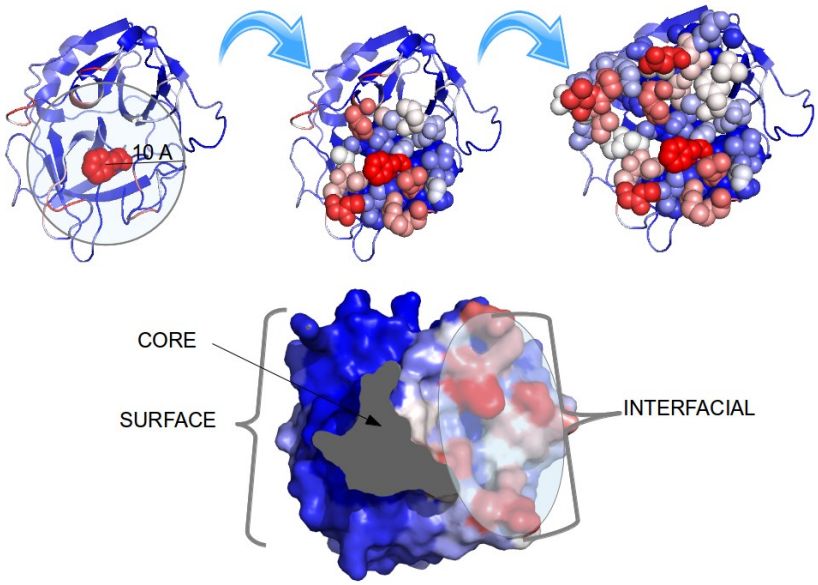



*Figure 5:* ***Scheme of prediction of the interface in a monomer using the hotspot prediction.***

After the docking procedure, we used pyDockNIP module to obtain hotspots of the interaction. The next step consisted in set a radius of 10 A around each hot-spot. We used the value of rASA to collect the surface residues around the hotspots thus forming the predicted interface

The protein-protein docking BM 4.0 (Hwang et al. 2010), was used to test the performance of the above described method to predict extended protein-protein interfaces. We processed all the 176 complexes in the benchmark with our docking-based interface prediction protocol, starting from the structures of the unbound proteins. The predicted extended interface patches were compared to the real interfaces, which were composed of those residues within 10 Å of the partner molecule in the complex structure. **Figure 5** shows a schematic representation of this procedure. Then sensitivity and precision of the method were computed as follows.

$$Sensitivity\,(S) = \frac{True\ Positives}{True\ positives + False\ negatives} \qquad \text{eq(7)}$$

$$Presicion\,(P) = \frac{True\ Positives}{True\ postives + False\ positives} \qquad \text{eq(8)}$$

## 3.11 Construction of the disease interaction networks

Genes associated with different diseases were obtained from the OMIM database (Amberger et al. 2009). For each gene, the UniProt codes and nsSNPs variants were obtained from the *humsavar.txt*

file (release 2014_06 of June 11th, 2014; downloaded from www.uniprot.org). The nsSNPs were classified, according to the *humsavar.txt* file, as: i) disease-associated, when the mutation is known to cause a disorder; ii) polymorphism, when the mutation is believed to be a neutral mutation; and iii) unclassified, when the mutation is detected in one or few patients, but showed low statistical significance due the limited size of the sample. This yielded a total of 112,205 disease-related nsSNPs associated to 3,440 disease phenotypes (based on OMIM ID), involving 2,355 distinct proteins (based on UniProt ID).

The network for each of the corresponding gene list was obtained using the Interactome3D server(Mosca, Céol, and Aloy 2013). The network for each phenotype showed all the possible partners which interacted with our query using UniProt codes. Table 2 displays the expansion and the coverage in each of the phenotypes extended networks used. When possible, the best structures or models for the UniProt code corresponding to a protein was downloaded using from Interactome3D database(Mosca, Céol, and Aloy 2013). We selected the best structures and models with the highest coverage and sequence identity from the proteins.dat file of the database. We downloaded the existing structures and models for the corresponding protein complexes from Interactome3D. If several structures or models were present for the same single protein, then we assigned the same UniProt code, but each part was used in separate docking experiments. In the same manner, the best structure or model, with the highest coverage and sequence identity, for the interaction was selected using the interactions.dat file of the database

*Table 2: Structural coverage of the disease-related protein interaction networks analyzed in this work*

| **Phenotype** | **OMIM Code** | **Associated Proteins** | **Protein Interaction networks** | | | | |
|---|---|---|---|---|---|---|---|
| | | Proteins[a] | Disease nsSNPs[b] | **Proteins** | Proteins w/ Structure[c] | **Interactions** | Interactions w/ Structure[c] |
| **HIGM5** | 608106 | 2 | 1 | 17 | 17 | 21 | 5 |
| **LHON** | 535000 | 6 | 21 | 23 | 22 | 33 | 10 |
| **CRC** | 114500 | 10 | 29 | 270 | 222 | 691 | 145 |
| **MCI** | 608446 | 11 | 0 | 193 | 162 | 582 | 154 |
| **HIV-1** | 609423 | 25 | 1 | 91 | 80 | 142 | 72 |
| **CMH** | 192600 | 7 | 174 | 198 | 162 | 531 | 100 |
| All six diseases[d] | | 61 | 226 | 729 | 610 | 1934 | 449 |

[a] OMIM
[b] humsavar
[c] Available structure or homology-based model
[d] union

. With this, we gathered 630 different proteins with structure or model that correspond to 1816 interactions which contained 2786 nsSNPs. We performed 9217 docking simulations on interactions without structure and 491 docking simulations on interactions with structure. After docking we performed the same classification of residues as with BM 4.0, but since multiple structures can be present for a single protein, to avoid ambiguous classifications for the same residue we used for classification the average of rASA for the residue in each of the structures.

Residues were classified as core, interface and non-interacting surface. Core residues were those with relative ASA < 0.1 (relative ASA is the ASA value for a given residue over the ASA reference value of the corresponding residue type). Then, exposed residues (relative ASA > 0.1) were classified as interface residues if any of their atoms are found within 10 Å from another atom from a partner protein. The remaining residues are classified as non-interface surface. When multiple structures exist for a single protein, to avoid ambiguous classifications for the same residue we used for classification the average of rASA for the residue in each of the structures.

## 3.12 Statistical analysis of nsSNPs on disease-associated protein interaction networks

In this work, we selected six disease phenotypes for which there is detailed structural information for most of the individual proteins within the network, but low structural coverage of the protein-protein interfaces: Hyper-IgM syndrome 5 (HIGM5); Leber hereditary optic neuropathy (LHON); Colorectal cancer (CRC); Susceptibility to myocardial infarction (MCI); Susceptibility to HIV type 1 (HIV-1); and CardioMyopathy Hypertrophic variants 1 to 15 (CMH).

All nsSNPs found in *humsavar.txt* for the protein structural interaction networks associated to the six selected disease phenotypes, were mapped to the corresponding protein structure. For this, the human sequences with all the variants were downloaded in a FASTA format from the UniProt web page (“The Universal Protein Resource (UniProt)” 2007). Then, the sequence and numbering of the PDB files in our dataset were extracted and aligned with the corresponding FASTA sequence when the numbering was incorrect or shifted.

The observed/expected (O/E) ratios for the distribution of nsSNPs in the above mentioned

protein regions (core, interface and non-interacting surface) were calculated as the observed fraction of nsSNPs found in each protein region over the fraction of nsSNPs expected by chance in each protein region. The latter was estimated from the fraction of total residues found in each protein region in all analyzed proteins.

The preference of a nsSNP for being at a given protein region *i* rather that at a region *j* was computed as an odds ratio (OR), as previously described (David et al. 2012):

$$OR_{ij} = \frac{\frac{P_i}{(1-P_i)}}{\frac{P_j}{(1-P_j)}} \qquad \text{eq(9)}$$

where $P_i$ is the probability of observing a nsSNP of a given type in protein region *i*, and is computed as:

$$P_i = \frac{n_i}{N_i} \qquad \text{eq(10)}$$

where $n_i$ is the number of nsSNPs of a given type observed at protein region *i*, and $N_i$ is the total number of residues at protein region *i* in all the analyzed proteins. A two-tailed P-value of less than 0.05 was considered indicative of the statistical significance of a preference for nsSNPs to be in one region over another compared to a random distribution based on the number of residues in the regions. Statistical analysis was performed using the statistical packages in R (3.1.1 version).

## 3.13 Identification of interface pathological mutations at RAS/MAPK cascade.

We used our interface prediction method to extend a previous study (Kiel and Serrano 2014) on 956 RASopathy and cancer missense mutations found in 15 genes of the RAS/MAPK pathway: PTPN11, SOS1, RASA1, NF1, KRAS, HRAS, NRAS, BRAF, RAF1, MAP2K2, MAP2K1, SPRED1, RIT1, SHOC2 and CBL. For the determination of possible pathways affected by the nsSNPs at the interface of the proteins, we used the GO annotation for the functional classification of genes provided by PANTHER database (Mi 2004).

## 3.14 Network graph and analysis

We used Cytoscape 3.4.0 (Shannon et al, 2003) to plot the different networks used in the thesis. This is an open source plotting program with a broad community of users that actively support systems biology analysis. For the enrichment of molecular function GO terms in the networks we used the Cytoscape plugging BINGO (Maere, Heymans, K. and Kuiper, 2005), with default parameters for a simple hyper-geometric test.

## 3.15 Interactome and core diseasome analysis with the combined expanded NIP strategy

We applied the same workflow we previously devised for the study of the six selected diseases to the high-confidence human interactome reported in 2014 by Rolland and coworkers (Rolland et. al 2014). As a complementary analysis, we also obtained the list of protein-protein interactions for the core diseasome reported in 2012 by Janjić and Pržulj (Janjić, and Pržulj. 2012). The core diseasome is a PPI network of key proteins involved in 561 GO terms related to pathologies that has been generated by computational analysis using a clustering technique known as k-Core decomposition of the *H. sapiens* PPI network data from BioGRID (Breitkreutz 2008) and HPRD (Keshava Prasad 2009) databases. We obtained all the structures or homology models for proteins and protein complexes in that network from Interactome3D. In total, merging the interactome and core diseasome, we analyzed a total of 4,254 different proteins that formed a total of 11,925 interactions. Among them, there is available complex structure or a straightforward homology model for only 2,226 interactions have available which involve a total of 2,039 proteins with structure or homology model. We performed protein-protein docking in 34,810 PDB pairs without a structure for their interaction. For testing purposes, we also performed docking on the 2,226 interactions common to the interactome and the core diseasome with structure or homology model. **Table 3** gives further details on the exact number of docking runs performed by each method in the human interactome and the core diseasome.

For this analysis, we used two docking protocols: pyDock and ZDOCK. The interface residue predictions were based on the NIP protocol extended by neighbor residues, as previously described (methods section 3.9). The original NIP analysis is done on the docking models generated by FTDock and scored by pyDock, whose performance when extended with neighbor residues was previously assessed (methods section 3.9). In addition, here we also used ZDOCK docking poses to

obtain the NIP values, with two different strategies: one using the ZDOCK docking poses as directly ranked by ZDOCK default scoring, and the other based on the ZDOCK poses as ranked by pyDock scoring function (PYDOCK_TOT). Thus, we generated three different NIP predictions: the original one based on FTDock rescored by PYDOCK_TOT, plus those based on ZDOCK with default scoring and on ZDOCK rescored by PYDOCK_TOT. Then, for each residue we kept the maximum of these three NIP values, from which the predicted extended interface was generated, using the neighbor residues as previously described. The performance of these new approaches for NIP calculation was assessed on the BM 4.0.

*Table 3: Number of docking runs performed by each method in the human interactome and the core diseasome networks*

| Docking method | | Human Interactome | Core Diseasome |
|---|---|---|---|
| pyDock | Protein complex with Structure or homology model | 2223 | 31 |
| pyDock | PDB pairs from Interactions w/o strcuture | 28932 | 5522 |
| ZDOCK | Protein complex wth Structure or homology model | 2220 | 31 |
| ZDOCK | PDB pairs from Interactions w/o strcuture | 28778 | 5522 |

The difference in docking runs between pyDock and ZDOCK comes from the availability of memory to perform the discretization of the surface of certain proteins into the corresponding grid. pyDock parameters are optimized for the use in parallel computing, this allows the use of shared memory from several processors, while ZDOCK only runs in single processor, thus the memory is limited. This is the reason why the geometry and size of certain proteins did not allow to perform a docking with ZDOCK.

“Extraordinary claims require extraordinary evidence.”

— Carl Sagan

# Chapter 4 Results

## 4.1 Performance of scoring functions in evaluating different protein-protein docking methods on the protein docking BM 4.0

We evaluated the performance of the 73 functions for the scoring of rigid-body docking poses generated for the BM4. **Figure 6A** shows the performance of the ten most successful functions ordered by top 10 success rate for FTDock, ZDOCK and SDOCK, respectively. In general, scoring functions provided better predictive rates when evaluating ZDOCK and SDOCK models. Interestingly, for all docking methods, there were always other scoring functions that performed better than its own in-built scoring method. The three scoring functions were found among the ten most successful ones for all docking methods were: AP_PISA(Viswanath, Ravikant, and Elber 2012), CP_TSC(Tobi and Bahar 2006), and CP_HLPL(Park and Levitt 1996; Pokarowski et al. 2005). The function CP_HLPL was originally developed for describing intramolecular contacts in protein structure modeling. The functions CP_TSC and AP_PISA were specifically designed for protein-protein docking using linear programing to train both functions. CP_TSC is a coarse-grain potential with three interaction sites per residue (side-chain centroid and N and O backbone atoms), which calculates the energy of interacting pairs with a two-step potential well. AP_PISA is an atomic potential which has a three-step potential between atom pairs, and was trained using side chain refined interfaces. These two potentials showed the best performance for the three docking methods, with AP_PISA being particularly successful in evaluating docking models generated by ZDOCK and SDOCK methods, when considering both the top 1 and the top 10 success rates. One of the possible reasons for the difference in performance of the three docking methods is the high variability in the total number of near-native solutions generated by each method (FTDock: 1,653; SDOCK: 18,700; ZDOCK: 37,709). This is an important factor that clearly can affect the capabilities of the scoring functions of discriminating near-native solutions from false positives.

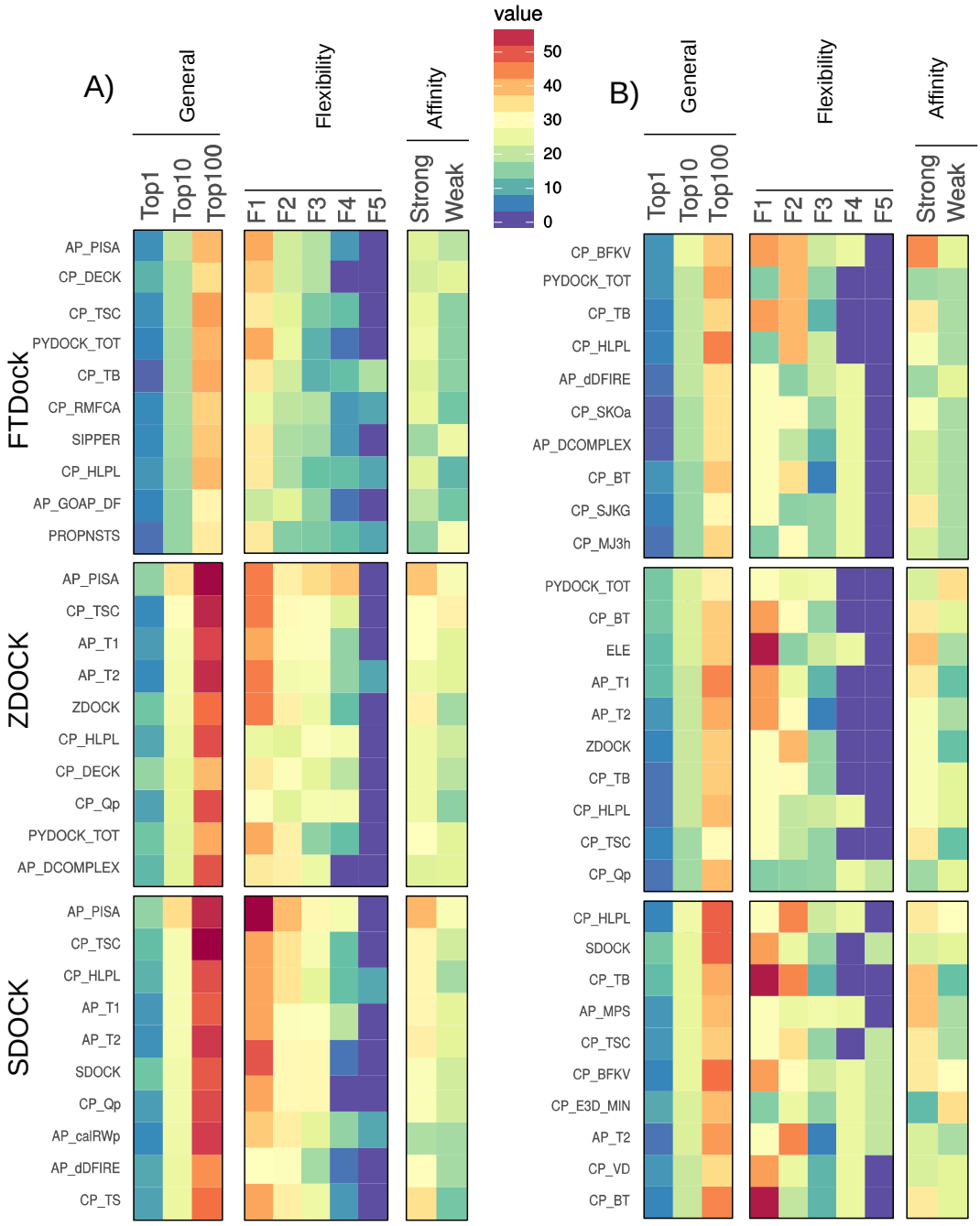



**Figure 6: Performance of scoring functions on (A) BM 4.0, and (B) BM 5.0.**

The first three columns show the success rates for the ten best performing scoring functions for each docking method, for the top 1, 10 and 100 predictions. Only the ten best performing scoring functions according to top 10 success rates are shown. Columns F1-F5 show success rates for the top 10 predictions according to conformational changes upon binding (F1: rigid; F2: low-flexible; F3: medium-flexible; F4: flexible, and F5: highly-flexible; see Methods). The two last columns show the success rates for the top 10 predictions according to binding affinity (see Methods).

## 4.2 Performance of scoring functions according to protein flexibility

The predictive success of rigid-body docking is known to depend strongly on the degree of conformational change that interacting proteins undergo upon binding(Carles Pons et al. 2010). We evaluated here whether this also applies to all scoring functions. For that, we classified the BM 4.0 cases according to the extent of unbound-to-bound conformational changes, based on the average interface RMSD (avgeIRMSD) for unbound receptor and ligand when superimposed onto the corresponding molecules in the complex structure, thus defining five categories: "rigid" (avgeIRMSD ≤ 0.5 Å), "low-flexible" (0.5 Å < avgeIRMSD ≤ 1 Å), "medium-flexible" (1 Å < avgeIRMSD ≤ 2 Å), "flexible" (2 Å < avgeIRMSD ≤ 3 Å), and "highly-flexible" (avgeIRMSD >3 Å).

**Figure 6A** shows the top 10 success rates for the above-analyzed scoring functions on models generated by each docking method, for cases classified according to unbound-to-bound conformational changes. In general, for each combination of scoring function and docking method, the best success rates are obtained for the rigid cases, as expected, and the performance decreases for the most flexible cases. However, there are interesting exceptions. For instance, AP_PISA on ZDOCK models provided better performance on the medium and flexible cases than on the low-flexible ones, and almost as good as on the rigid ones. Similarly, the performance of CP_HLPL on ZDOCK was independent of the flexibility category. Interestingly, a few combinations of scoring functions and docking methods identified acceptable docking models within the top 10 decoys for the highly-flexibly cases. These are extremely challenging for rigid-body docking prediction, so the fact that some scoring functions can predict some of them is in principle encouraging. However, due to the smaller number of flexible cases in the benchmark these differences are not statistically significant (Wilcoxon signed rank test FTDock p-value = 0.333, ZDOCK p-value = 0.667, SDOCK p-value =0.333). Only 6% (11 cases) in the BM4 corresponds to the highly-flexible category which contains the monomers that undergo the biggest conformational change upon binding to form a complex. In general, the performance of the different scoring functions on the rigid cases shows more consistency, while that on the most flexible cases shows more variability, which suggests possible random effects on the latter due to lower signal-to-noise ratios

## 4.3 Performance of scoring functions according to binding affinity

The predictive performance of rigid-body docking also strongly depends on the binding affinity of the complex (Vajda 2005). High-affinity complexes are in general predicted with higher accuracy.

We have explored here to what extent the performance of the different scoring functions depends on the affinity of the complexes. For that, we gathered the experimental binding affinity data from the affinity benchmark(Kastritis et al. 2011) for 89 cases of the protein-protein docking BM 4.0. Then we classified these cases into two categories according to their binding ΔG value: “Strong” (ΔG ≤ -12 Kcal/mol, and “Weak” (ΔG > -12 Kcal/mol ).

**Figure 6A** provides the success rates for the top 10 predictions of the previously analyzed scoring functions for the different docking methods on the benchmark BM 4.0 cases as classified by binding affinity. In general, predictive performance on the strong affinity cases is better than on the weak affinity cases. However, there are some exceptions, being the most notable ones the SIPPER SIPPER (C. Pons et al. 2011), and PROPNSTS (Liu et al. 2004)) functions when evaluating FTDock models, which yielded much better predictions for the weak affinity cases. Interestingly, these two scoring functions are based on the same residue potentials derived from protein-protein complex structures. It seems that they are able to capture the binding energy determinants of weak complexes better than other atomistic potentials.

## 4.4 Performance of scoring functions on the CAPRI scorer set benchmark

We evaluated the performance of the 73 scoring functions on the scorers set benchmark, which is formed by 15 targets from the CAPRI experiment (Lensink and Wodak 2014), for which a range of docking models was blindly generated by a variety of docking methods (see Chapter Methods section 3.2). **Figure 7A** shows the predictive rates for the best 30 scoring functions in this benchmark according to the top 10 success rate.

We found here some of the most successful scoring functions overlap with those that have performed well on BM 4.0, such as AP_T1 (Tobi 2010), AP_T2 (Tobi 2010), CP_DECK (Liu and Vakser 2011), CP_TB (Tobi 2010), CP_TSC and AP_PISA. Interestingly, the best success rates for the top 100 predictions were obtained by coarse-grain potentials, in general. Perhaps coarse-grained potentials are providing a more balanced score that is more adequate to the heterogeneity of docking models generated by a large variety of docking methods in this scorers set benchmark from CAPRI. The most successful function for the top 100 predictions is CP_TB, a scoring function designed for docking, which was among the most successful ones with FTDock on the BM 4.0.

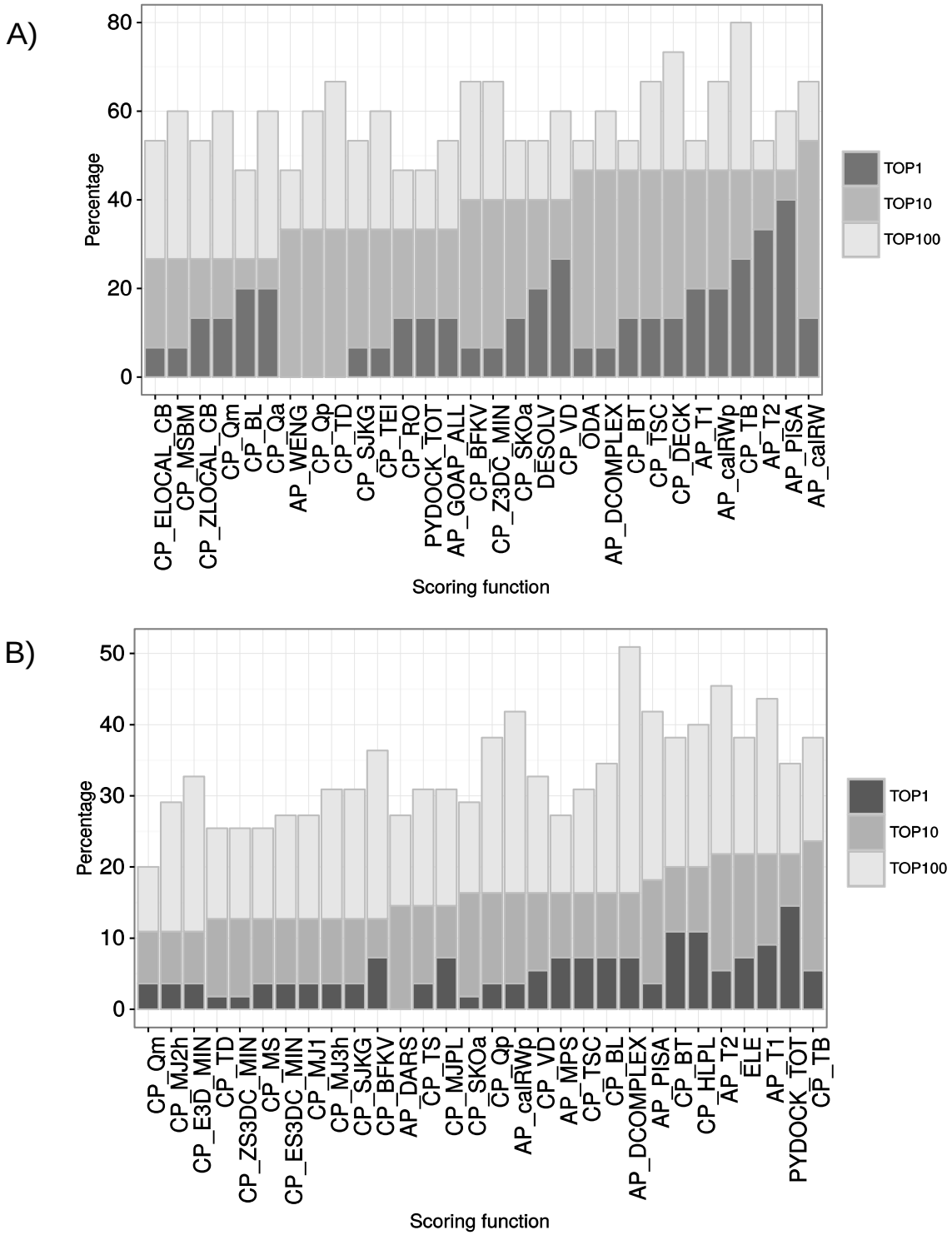



*Figure 7:* **Performance of scoring functions on different docking sets.**

A) Success rates on the CAPRI score set benchmark. (B) Success rates on the docking sets obtained by merging the results from the three docking methods. Only the 30 best performing scoring functions are shown.

## 4.5 Performance of the scoring functions with different docking methods on protein docking BM 5.0 update

In order to confirm the previous findings, we also evaluated the performance of the different scoring functions and docking methods on the recently available protein docking BM 5.0 update (Vreven et al. 2015), formed by cases that were not present in BM 4.0. This analysis provided unexpected and interesting results. **Figure 6B** shows the success rate of the ten most successful functions (when considering the top 10 predictions) for FTDock, ZDOCK and SDOCK, on the BM 5.0 update. We can observe that the best scoring functions now are different from the best scoring functions of BM 4.0 **(Figure 6A)**, especially for ZDOCK and SDOCK. One of the most striking differences is AP_PISA, which shows much lower performance than on BM 4.0, and may be indicative of overfitting to the BM 4.0 complexes during training. This may also explain some of the other differences observed**.**

**Supplementary Figure 1** and **2**shows the performance on the BM 5.0 update of the best-performing functions resulting of the previous BM 4.0 analysis. he predictive rates of these functions for the BM 5.0 update are much lower than those observed for the BM 4.0 cases. In addition, there are now less differences in the best predictive rates for the different docking methods. Indeed, now the evaluation of ZDOCK and SDOCK docking models does not show better success rates than FTDock as is the case for the BM4.0. The performance for the scoring functions on the FTDock models, with similar success rates on both BM 4.0 and 5.0, is more consistent than that of ZDOCK and SDOCK. It seems that the performance obtained for some scoring functions on BM 4.0 with ZDOCK and SDOCK were excessively high. One reason could be that these scoring functions might have been overtrained on cases from BM 4.0, using ZDOCK and SDOCK methods to generate docking decoys, another explanation could be that SDOCK is technically similar to ZDOCK.

The BM 5.0 update provides a set of cases that were not used for training, since it does not include complexes from previous benchmark sets. A key question is whether the best-performing scoring functions for BM 5.0 update represent bona fide success rates for docking in general or they appear good only for this particular set of cases for another reason. The fact that CP_HLPL and CP_TB are found among the best-performing scoring functions with the three docking methods on BM 5.0 update, suggests that their good performance on BM 4.0 was not due to overtraining, and

therefore they could be of more general applicability for new cases. Indeed, CP_HLPL, which used with SDOCK provided the best top 10 success rate among all functions (25%), was originally developed from intramolecular contacts for modeling protein monomers. On the other side, CP_TB was developed for docking but trained in a composite set of representative transient complexes. This knowledge-based potential was designed to tolerate small changes in side chain orientations, which may contribute to its avoidance of overtraining.

## 4.6 Performance of scoring functions according to binding affinity and flexibility in BM 5.0 update cases

We have analyzed the results of the cases in BM 5.0 update as classified according to unbound-to-bound conformational flexibility **(Figure 6B)**. Several functions (CP_TB with FTDock and SDOCK models; CP_BT, CP_BFKV and CP_SKOa with FTDock, etc.) can provide similarly good performance for rigid and low-flexible cases.

We also analyzed the results for the 35 cases of the BM 5.0 update for which there is experimental binding affinity available (Vreven et al. 2015). These cases were classified as strong or weak, according to their experimental binding affinity **(Figure 6C)**. The affinity-dependent performance of some of the scoring functions varies according to the docking method. For instance CP_HLPL shows no dependence on affinity for SDOCK, but strong dependence for FTDock. The performance of some of the functions for the strong binders in the BM 5.0 update is better than those in the BM 4.0, perhaps due to the fact that the BM 5.0 has fewer cases with affinity information.

## 4.7 Scoring performance on models merged from different docking methods

We merged all docking models generated by the three docking methods into a single decoy set, and evaluated the performance of each scoring function on this heterogeneous pool. We have evaluated the performance of each scoring function on this heterogeneous pool of docking solutions. Figure 7B hows the performance for the best 30 scoring functions on this set ordered by top 10 success rates. In general, the success rates for the best performing functions were lower than those obtained with the individual methods. For instance, the best performing scoring function on the merged pool of docking models is CP_TB, with 24% success rate for the top 10 predictions, while for the

individual methods, CP_HLPL with SDOCK, and CP_BFKV (Feng, Kloczkowski, and Jernigan 2010) with FTDock yielded higher success rates (over 25%). Surprisingly, these scoring functions yielded much lower success rates on the large docking set (20% and 12%, respectively)

This shows that some scoring functions are particularly efficient for a specific docking method, so it seems more reasonable to use each docking method only with the scoring functions that have shown the best performance on such method. A different question is which scoring function to use when we do not know which docking method was used to build each docking model. In this case, a good scoring function that might work for a particular method (i.e. CP_BFKV on SDOCK and FTDock) might give worse predictive rates in other docking method (i.e. CP_BFKV on ZDOCK). In this situation, it would be better to choose some more general scoring function that could provide good success rates in all methods (i.e. CP_TB or PYDOCK_TOT). This could be relevant in the CAPRI scorers experiment, for instance, in which a variety of docking models need to be scored, but there is no information given on how they were generated

## 4.8 Performance of combined scoring functions

We next explored whether the combination of scoring functions might improve the predictive rates. First, we identified pairs of scoring functions that provided successful results in complementary subsets of complexes. The first metric we used to do this is the size of the combined set of complexes for which an acceptable or better solution was found in the top 10 by either of the scoring functions (union cardinality). The second metric was similar, but excluding the complexes that are identified by both functions (symmetric difference cardinality). These measures were chosen to give an indication of how both scoring functions bolster each other, and therefore, this could be used as an estimation of the potential synergistic effect of the two functions when combined. **Figure 8** shows the cardinality values (for top 10 predictions) for the combinations of the ten functions with the greatest union values when paired, for each of the docking methods on the BM 5.0 update. **Supplementary Figures 3-8** shows these values for all pairs of scoring functions. We can observe that some pairs of scoring functions are highly complementary, since they are able to capture near-native solutions on non-overlapping sets of complexes (e.g. PYDOCK_TOT/CP_BFKV with FTDock; PYDOCK_TOT/AP_T2 with ZDOCK; AP_MPS/SDOCK or AP_MPS/CP_RMFCEN1 with SDOCK).

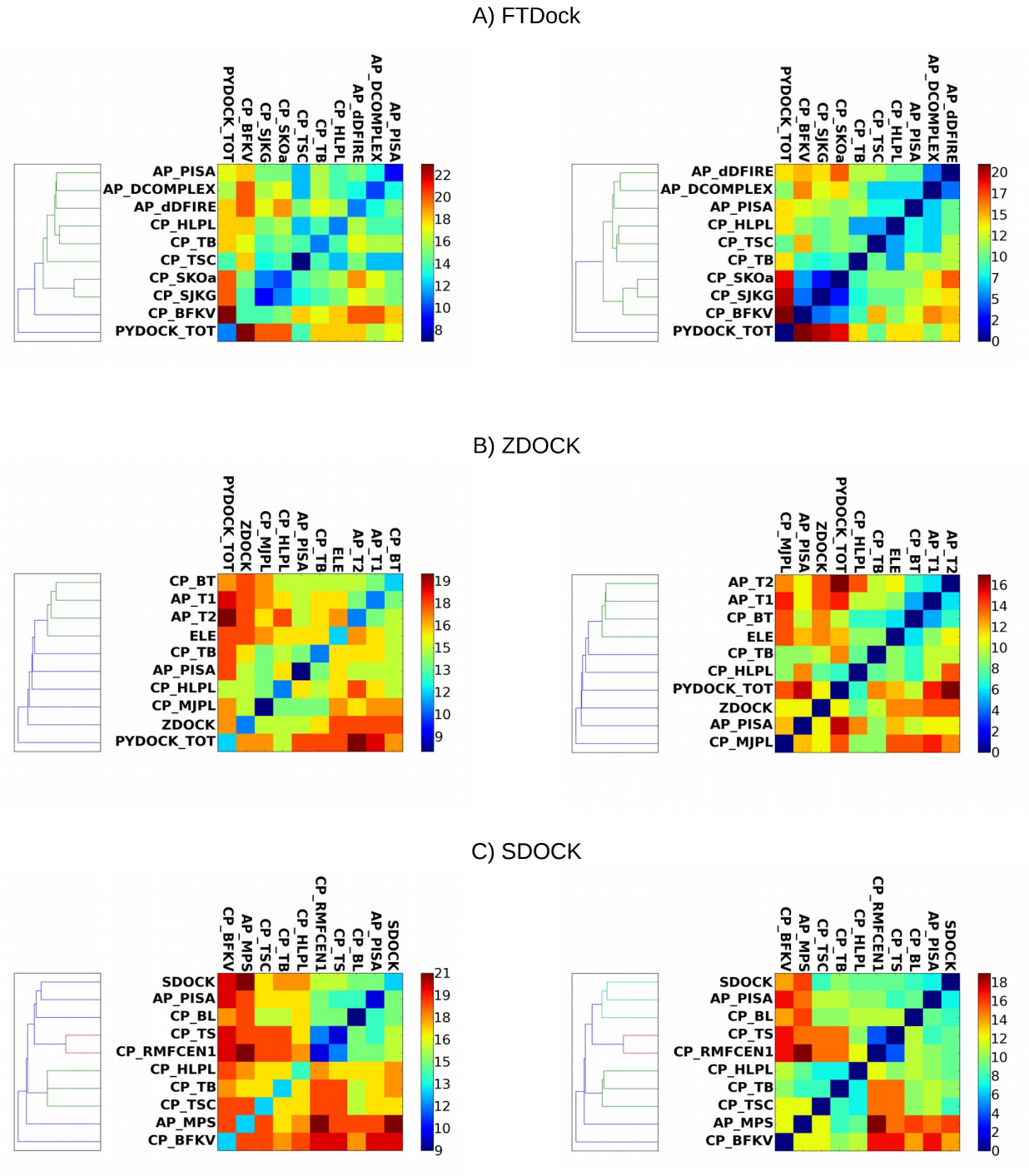



*Figure 8:* ***Cardinality analysis on the different FFT methods using BM 5.0***

The heat-maps show the union (right) and symmetric difference(left) values for pair combinations of the ten scoring functions that provided the highest union values (top 10 predictions) for each docking method on BM 5.0 update, with functions grouped using single linkage clustering..The heatmaps show the scoring function pairs with highest cardinalities for top10 hits, with functions grouped using single linkage clustering.

From the above analysis, one could estimate the most favorable pairs of scoring functions, i.e. those ones that when combined should yield improved success rates. Therefore, we tested the predictive power of the best pairs of functions derived from the cardinality analysis. For this, we normalized the energy values obtained from each pair of functions and converted them into z-scores. Then we added these values without weighting and used them to re-rank all the generated decoys for each case. **Figure 9** shows the predictive rates (on BM 5.0 update) for the combinations of the ten scoring functions that provided the largest union values (for top 10 predictions) on the BM 5.0 update. Some combinations yielded >30% success rates for FTDock models (as compared with 20-25% for the individual scoring functions). However, in the case of ZDOCK and SDOCK docking methods, success rates of the best combined scoring functions did not improve the individual ones. This small improvement in the success rates for a few combinations of scoring functions is not sufficient to guarantee that this strategy could be of general applicability to a new set of cases, and requires further investigation.

For some pairs of scoring functions, the cardinality analysis did not reflect well the success rate values after rescoring with the combined functions. For instance, with FTDock, the best union was found for PYDOCK_TOT/CP_BFKV, providing near-native models for 42% of the cases within the top 10 predictions, but the combined functions have a top 10 success rate of 27%. Individually, PYDOCK_TOT has a top 10 success rate of 20% and CP_BFKV of 26%. For some reason PYDOCK_TOT seems to contribute little at the combined success rate in the top 10 predictions in spite of the observed high cardinality. On the other hand, the best top 10 success rates after rescoring with the combined functions is provided by CP_SJKG/AP_dDFIRE (33%), while individually, CP_SJKG and AP_dDFIRE have much lower success rates (16% and 20%, respectively). For this pair, the union was not among the best values of all cases, so cardinality analysis was not able to foresee the strong synergy shown by the combination of these two scoring functions.

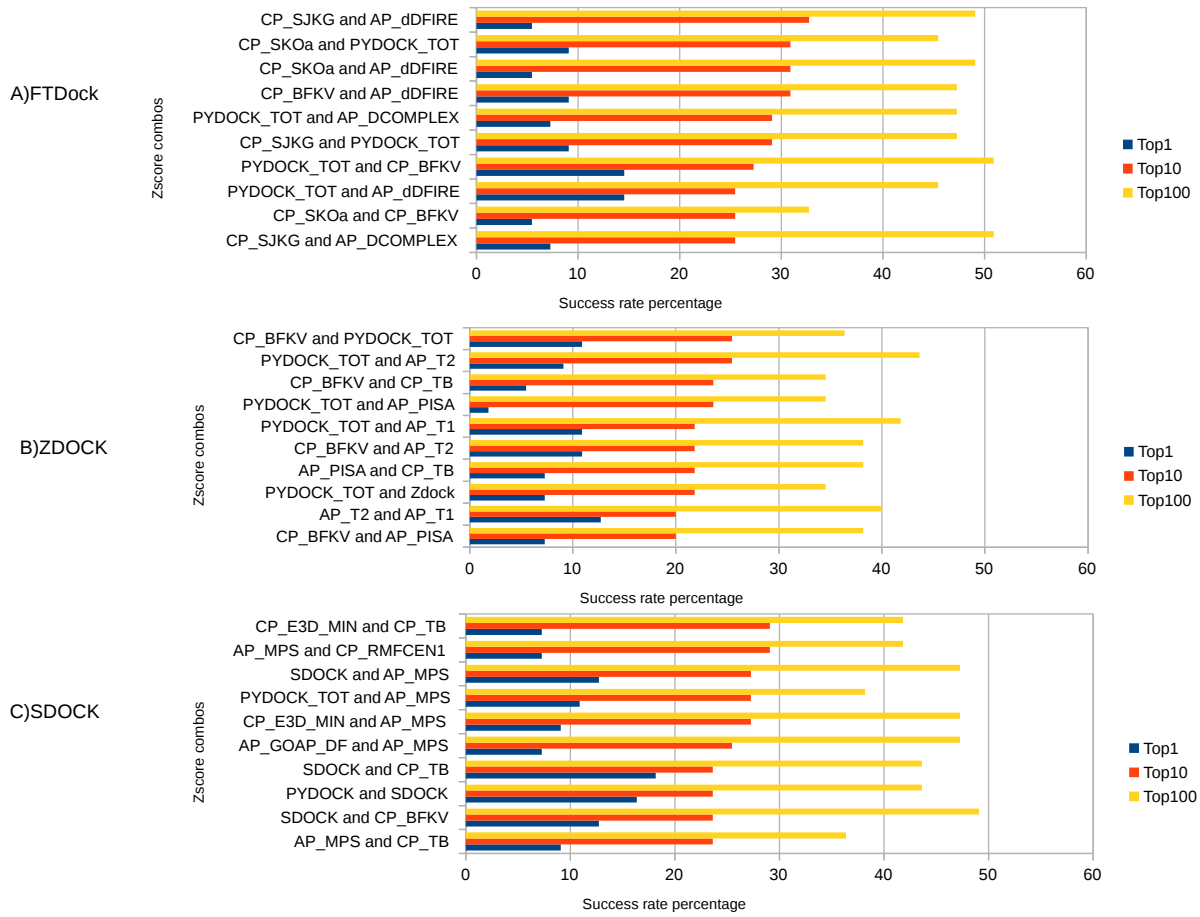



*Figure 9:* ***Success rates on BM 5.0 for pair combinations of scoring functions using z-scores.***

Performance on BM 5.0 of scoring function pairs formed by unweighted combination based on z-scores, using the ten scoring functions that provided the best union values (for top 10 predictions) on BM 5.0 for each docking method. The ten pairs of scoring functions with the best top 10 success rates are shown for each docking method: A) FTDock, B) ZDOCK, and C) SDOCK.

So far, we selected the top scoring functions for each docking method in BM 5.0 update and evaluated its performance in the BM 5.0 itself. To make a blind test, we selected the ten scoring functions with the best top 10 success rates from BM 4.0, and computed their cardinalities on BM 5.0 update (**Figure 10**). With FTDock the best cardinalities are found for combined pairs involving PYDOCK_TOT, being the highest ones the combinations with CP_HLPL and CP_TB. With ZDOCK there are many combinations that give a high cardinality, such the combination of PYDOCK_TOT with AP_T1/2 or AP_PISA.

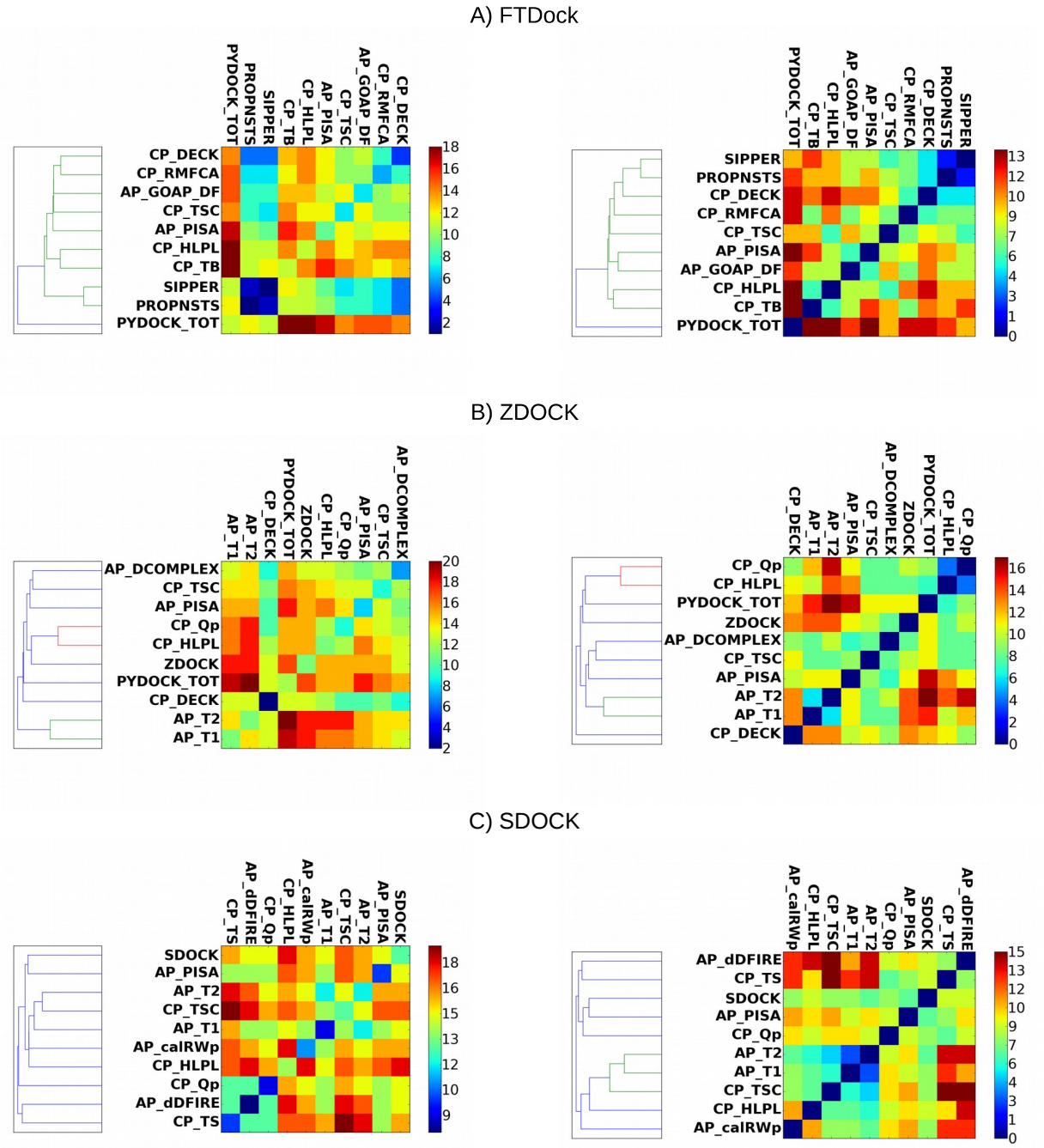



*Figure 10:* ***Cardinality analysis on the different FFT methods using BM 4.0 top scoring functions evaluating the BM 5.0.***

*See figure 5 for details*

**Figure 11** shows the top 10 success rates on BM 5.0 for the best scoring function pairs formed by unweighted combination based on z-scores, using the ten scoring functions that showed better performance from BM 4.0. We found two pairs of combinations that reached a success rate above 30% within the top 10 predictions: the pair PYDOCK_TOT/CP_HLPL with FTDock (31%),

and the pair AP_PISA/CP_HLPL with SDOCK (31%). Thus, combined pairs formed by scoring functions selected on the basis of BM 4.0 yielded success rates on BM 5.0 as high as those obtained when the combined pairs were formed by functions selected among the best ones in BM 5.0 update, which suggests little or no overfitting.

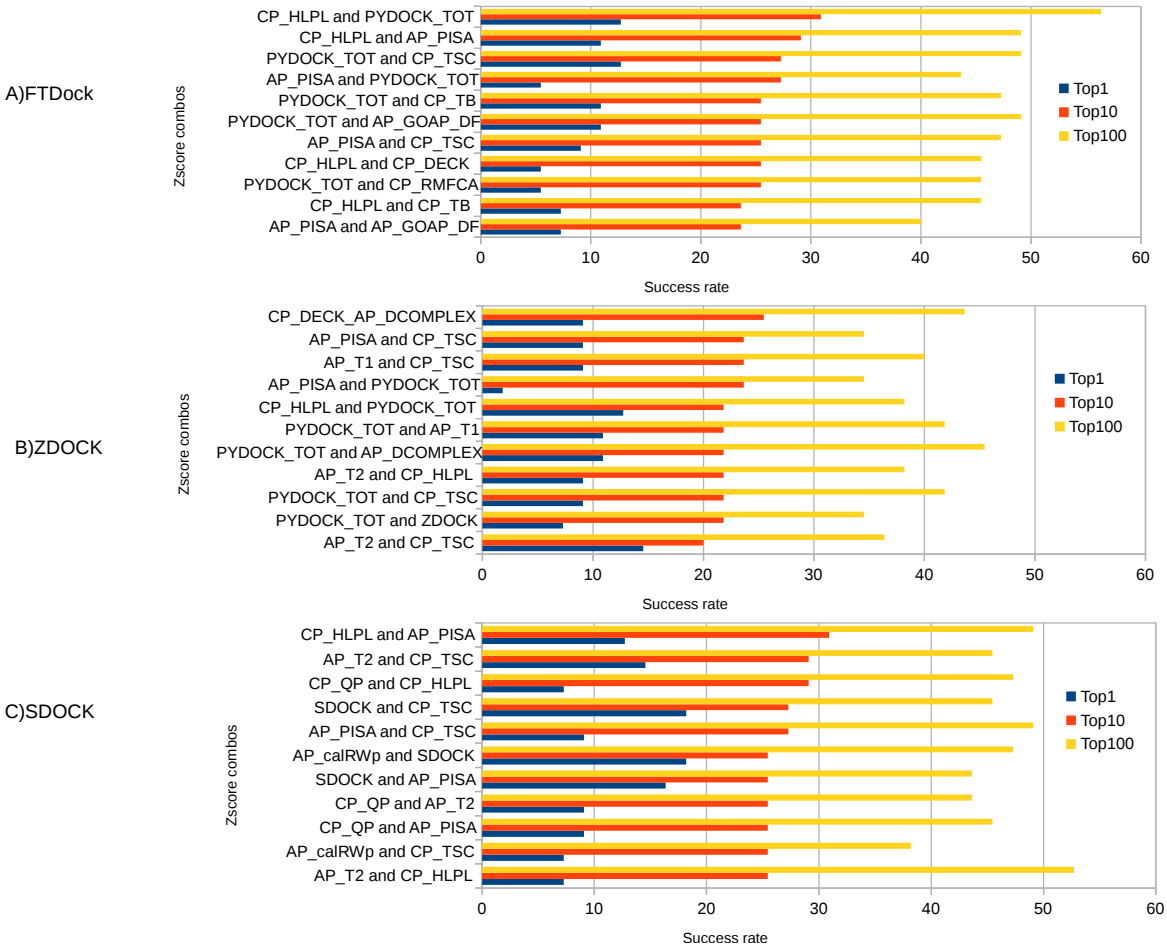



*Figure 11:* **Success rates on BM 5.0 for pair combinations of the best-performing scoring functions from BM 4.0.**

Performance on BM 5.0 of scoring function pairs formed by unweighted combination based on z-scores, using the ten most successful (top 10 predictions) scoring functions from BM 4.0, for each docking method: A) FTDock, B) ZDOCK, and C) SDOCK.

Overall, this is not a considerable increase in the success rate. To extend the number of existing near-native solutions and possibly improve the scoring performance, a heterogeneous pool of decoys could be created from the three docking methods and the best scoring functions for each docking method. In fact, a researcher is not limited to use only one docking method, e.g. the

complementarity of ZDOCK and FTDock both using the PYDOCK_TOT scoring was used to help to model yeast interactome.[39] In this line, we aimed to combine the pairs of scoring functions that performed well on each set of docking decoys generated by FTDock, ZDOCK and SDOCK, and tried to evaluate whether they would improve the predictive results. For this, we built scoring function pairs formed by unweighed combinations based on z-scores, using the ten scoring functions that provided the best top 10 success rates for each docking method in BM 4.0. With them, we built triplets of combinations formed by one pair of scoring functions from each docking method, and computed the union cardinality (for the top 10 predictions) for each triplet on BM 5.0. **Supplementary Table 2**. shows the combined triplets with the 50 best union cardinalities. The best triplet combinations generated by this strategy captured 30 cases (55%), considerably more than the 18 cases (33%) predicted by the best-performing pairs of scoring functions (CP_SJKG and AP_dDFIRE with FTDock) from the cardinality analysis carried out with the individual docking methods. According to these results, the use of triplet combinations of the best pairs of functions for each method seemed to anticipate a large improvement in success rates. To confirm this, we used the best scoring function pairs for each method (according to BM 4.0), and computed the success rates of the triplet combinations of function pairs / docking methods on BM 5.0 (**Figure 12**). The best triplet combination is formed by PYDOCK_TOT and CP_HLPL with FTDock, AP_T2 and AP_PISA with ZDOCK, and AP_calRWp and SDOCK scoring function with SDOCK (38% success rate). However, despite the expectances, this is not much better than the best performance we found for a pair of scoring functions (CP_SJKG/AP_dDFIRE with FTDOCK; top 10 success rate 33% on BM 5.0).

The combination of scoring functions performed here was based on a direct addition of the normalized functions. There was no attempt to improve the combination of values, by optimization of parameters, multi-parametric fitting, etc. However, due to the process of selection of scoring functions, there could be a possible bias towards the best-performing functions on the BM 5.0. The use of more sophisticated approaches to combine the scoring functions could yield better predictive rates, but such analysis should be done with caution, to minimize the risk of overfitting, for instance by putting feature selection within an outer cross-validation wrapper.

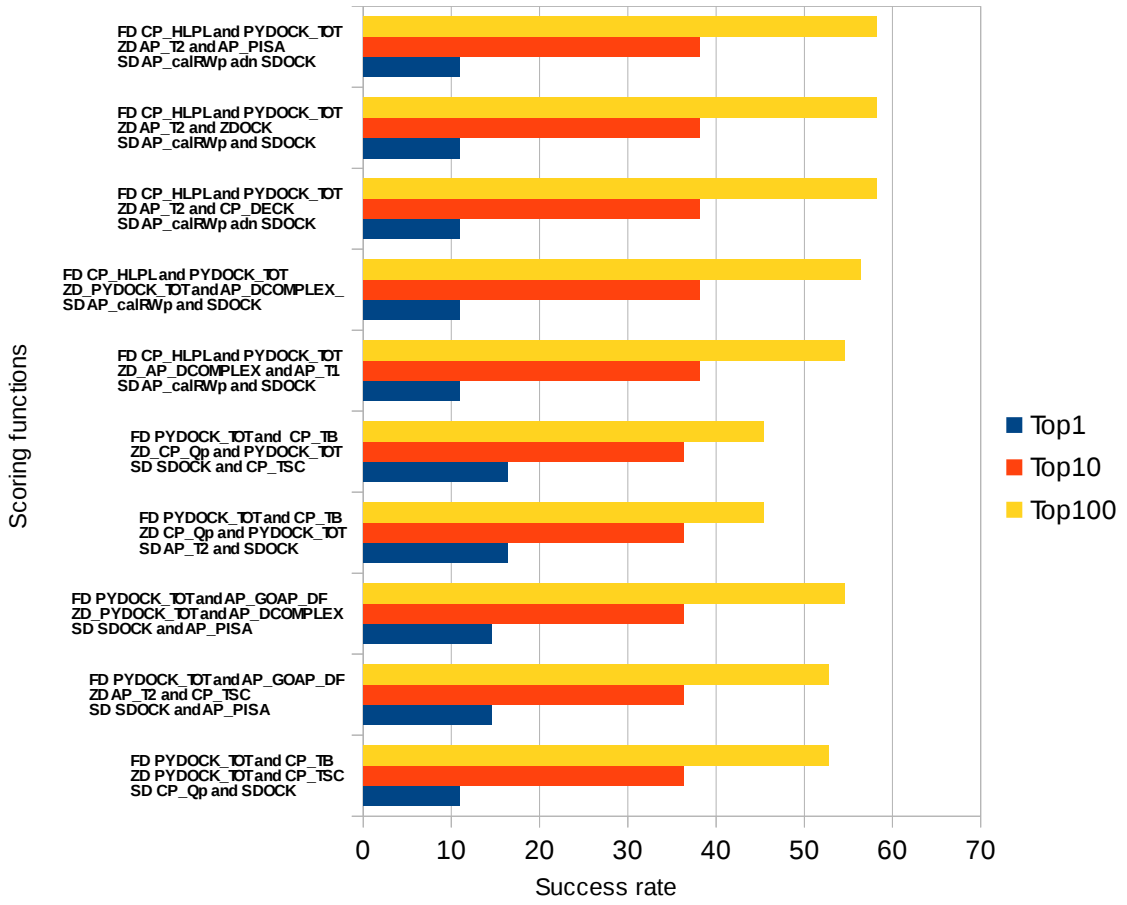



*Figure 12:* **Success rates on BM 5.0 for triplet combinations of the best performing scoring functions and docking methods from BM 4.0.**

Performance on BM 5.0 of the triplets formed by unweighted combination of scoring functions (z-scores) with each of the docking methods, using the ten most successful (top 10 predictions) scoring functions from BM 4.0 for each docking method. The origin of each scoring function pair is indicated at the beginning of each line as follows: FD from FTDock, ZD from ZDOCK, and SD from SDOCK.

## 4.9 Consensus ranking of protein-docking decoys

So far, the combination of scoring functions performed here was based on a direct addition of the normalized functions. There was no attempt to improve the combination of values, by optimization of parameters, multi-parametric fitting, etc. However, due process of selection of scoring functions there is a possible bias towards the best-performing functions on the BM 5.0. The use of more sophisticated approaches to combine the scoring functions could yield better predictive rates, but such analysis should be done with caution, to minimize the risk of overfitting, for instance by

putting feature selection within an outer cross-validation wrapper.

In order to attempt the combination of many scoring functions we decided to explore the use of machine learning algorithms to improve the ranking of the different methods. The decoys characterized by a large selection of scoring function from the previous study were used as descriptors or features, that were ordered by an ensemble of ranking support vector machines (R-SVMs) (Joachims 2005). A consensus ranking is calculated by combining the R-SVMs using the Schulze voting method. The method was applied independently to decoy structures from four state of the art docking programs, ZDOCK, SDOCK, to our docking protocol pyDock (not to confuse with PYDOCK_TOT which is the scoring function) and a non-FFT method SwarmDock (I. H. Moal and Bates 2010). Our collaborator I.H Moal obtained the docking decoys from SwarmDock and evaluated them with the same scoring functions. Swarmdock is regarded as one of the best protein-protein programs in the CAPRI contest, and its sampling is based on the Swarm Particle Optimization (SPO) algorithm. We used this program as an important point of comparison for the enhancement of the FFT based methods and the precedent of the work done with the diverse scoring functions.

To validate, we trained the models using the protein-protein docking BM 4.,0 and evaluated the ability to retrieve near-native solutions using the new complexes added in the BM 5.0 as an external validation set (**Figure. 13**). Of the complexes for which a near-native solution could be found, a near-native structure was identified as the top-ranked solution in 12-22% of the interactions prior to re-ranking, which increased to 16-44% using our approach. Similarly, retrieval in the top 10 increases from 33-51%, to 50-67%, and top 100 improves from 70-90% to 91-100%, indicating that sampling becomes the limiting factor in obtaining a top 100 near-native solution within our scoring scheme.

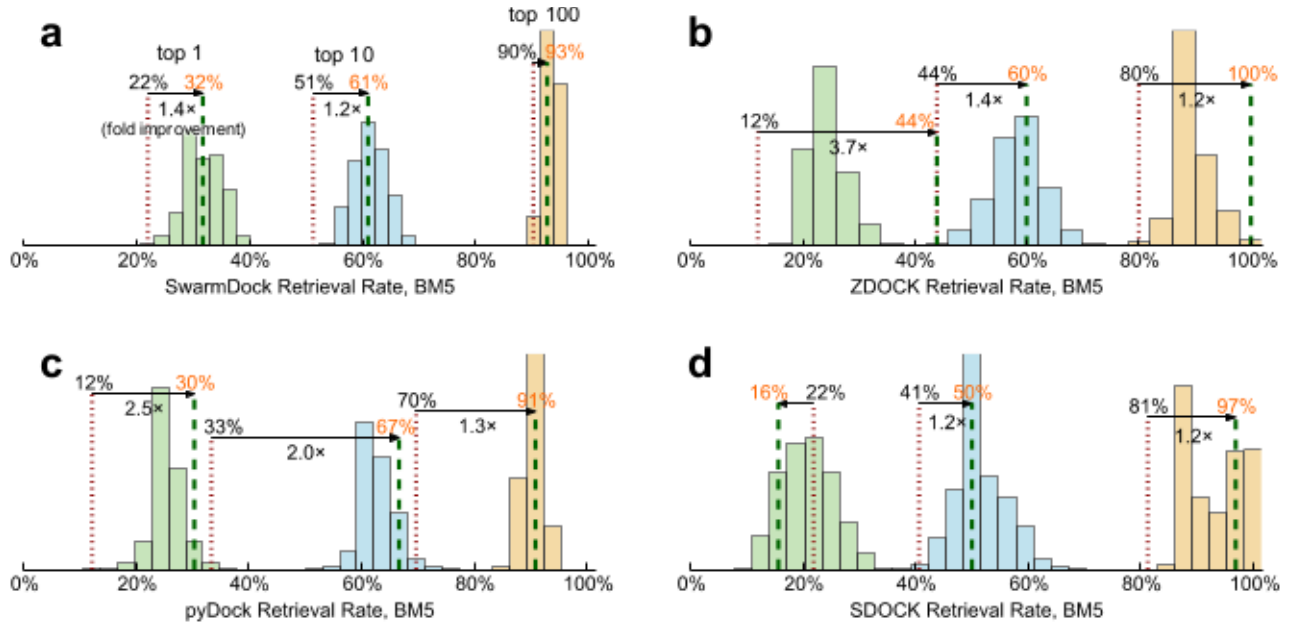



***Figure 13: Retrieval rate of the different methods on the BM 5.0.***

Panels show the top 1, top 10 and top 100 retrieval rates for the original (red dots) and consensus (green dashes) rankings, as well as the distributions for the ensembles of support vector machines when applied to the new complexes in the BM 5.0 as external test set

We globally boosted the success rates of each docking procedure on the BM 5.0 (**Figure 13 and Supplementary Table 3**) as following: top1 ranking to 24%, top10 ranking to 45 % and top 100 ranking up to 69%.We also applied the method to the original complexes in the docking BM 4.0 using multiple leave-many-out cross-validations **(Figure 14)**. A quarter of the complexes were left out at random from the training set for each of the R-SVM models, and for each complex, the Schulze re-ranking only combined the models for which the complex was omitted from the training.

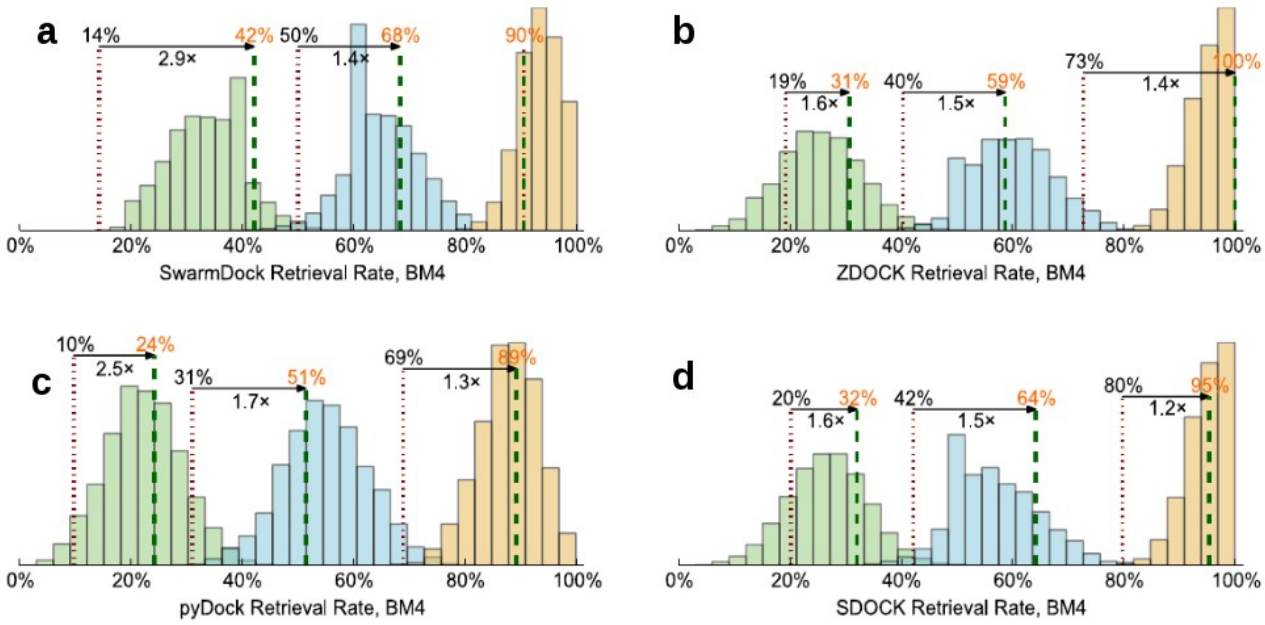



*Figure 14:* ***Retrival rate of the different methods in the BM 4.0.***

As in the previous figure *p*anels show the top 1, top 10 and top 100 retrieval rates for the original (red dots) and consensus (green dashes) rankings, as well as the distributions for the ensembles of support vector machines when applied to the new complexes in the BM 4.0 as external test set

We see improvements of 10-20% to 24-42%, 31-50% to 51-68% and 69-90% to 89-100% respectively for the top 1/10/100 retrieval rates. For SwarmDock, this corresponds to top 1/10/100 success rates of 30%, 49%, and 65% respectively, in the case of pyDock 14%,30%,52%, for ZDOCK 20%,38%,65%, and SDOCK 20%,40%,59%, when considering all 176 complex in the BM 4.0 typically performed (**Figure 15-right and Supplementary Table 3**). On both benchmarks, a large improvement can be attributed to the R-SVMs which, when combined using Schulze ranking, typically performing as good as or better than the average R-SVM model on its own. For all four docking protocols, the method yields a significantly better ranking of the top-ranked near-native solution ($p << 0.01$, Wilcoxon signed-rank test), and large improvements compared to other methods reported in the literature (**Figure 15 and Supplementary Table 3**).

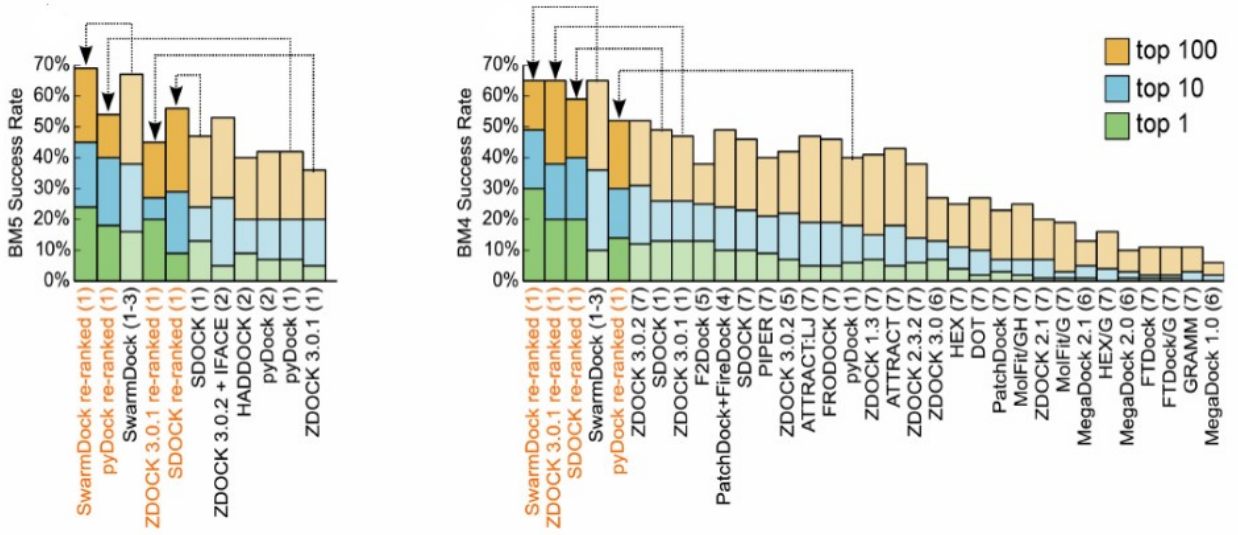



*Figure 15:* ***Comparison of success rates: machine learning and democratic ranking versus other docking methods.***

Panels show the top 1 and top 10 success rates for the whole docking pipeline for the 55 new BM 5.0 complexes (left)and the 176 BM 4.0 complexes (right), using data from this study(1), as original rankings (lighter colors) or using either the BM 50 complexes as external test set or multiple leave-many-out cross-validation with the BM4 (dark colors), and data reported in Vreven et al.(2) , Torchala et al.(3) , Schneidman-Duhovny et al.(4), Chowdhury et al.(5), Ohue et al.(6) and Huang (7).

## 4.10 Structural analysis of pathological mutations on protein interaction networks

After all the analysis on the scoring of docking poses described in the previous sections of this thesis, we next aimed to explore whether docking-based computational approaches could help characterizing disease-related mutations in PPIs at interactomic scale, where the majority of protein-protein interfaces have no structural data. For this purpose, we focused our analysis on the protein-protein interaction networks of six disease phenotypes for which there is detailed structural information for most of the individual proteins within the network, but low structural coverage of the protein-protein interfaces. **Table 2** (see Chapter Methods section 3.11) shows the number of proteins associated to each disease according to OMIM, as well as the number of proteins and complexes forming the first-layer interaction network and their structural coverage.

We first analyzed the location of the known nsSNPs within the protein interaction networks of the six analyzed diseases, considering only those protein-protein interactions that had available structure (or a reliable homology model). This structural dataset was formed by 449 protein-protein complexes that had available structure (or a reliable homology-based model), and involved 353 proteins with available structure (experimental or modelled). We found that 258 of these proteins had at least one annotated nsSNP (**Table 2**). The entire set comprised a total of 1,624 nsSNPs that could be structurally characterized, of which 832 were related to a disease (not necessarily any of the originally analyzed six diseases), 499 were classified as polymorphisms, and 293 were unclassified. Among the structurally mapped disease nsSNPs, 48% are buried, 22% are located at a protein-protein interface, and 30% are found at a non-interacting surface (**Figure 16A**). We can compare these numbers with the values expected by chance for buried, interface and non-interface residues (29%, 31% and 40%, respectively), as estimated from the residue composition of the studied proteins (see Chapter Methods section 3.12). Thus, the observed/expected (O/E) ratios for buried, interface and non-interface disease nsSNPs are 1.68, 0.70 and 0.75, respectively. The disease nsSNPs are located in buried positions clearly more often than expected by random, which has already been observed in previous studies (David and Sternberg 2015; David et al. 2012). However, the O/E value for the interface disease nsSNPs obtained here (0.70) is clearly below that reported in previous studies on a large interaction data set (0.96 (David and Sternberg 2015); an earlier studied found this value to be 1.20, but in that case interface residues were defined exclusively based on distance criteria and could include some buried residues (David et al. 2012)). More interesting is to analyze the preference of a disease nsSNP for being at a protein-protein interface rather that at a non-interacting surface, computed as an odds ratio (OR) (see Chapter Methods section 3.12). Here, we found that disease nsSNPs had similar probability of occurring at protein interfaces than at non-interacting surfaces (OR 0.94). Again, this value is lower than that previously reported on a large interaction dataset, in which they found a clear preference of disease nsSNPs to be at interface regions rather than non-interacting surfaces (OR 1.35 (David and Sternberg 2015)). The lower preferences found here for the disease nsSNPs to be located at protein-protein interfaces can be explained by the low structural coverage of the protein interactions in the six diseases studied here (which were indeed selected because they had high structural coverage for the individual proteins but low structural coverage on the protein-protein complexes). This shows that the lack of structural data on protein-protein complexes might underestimate the role of many disease nsSNPs involved in protein interactions and can lead to poor characterization of the effect of these mutations in the

network topology.

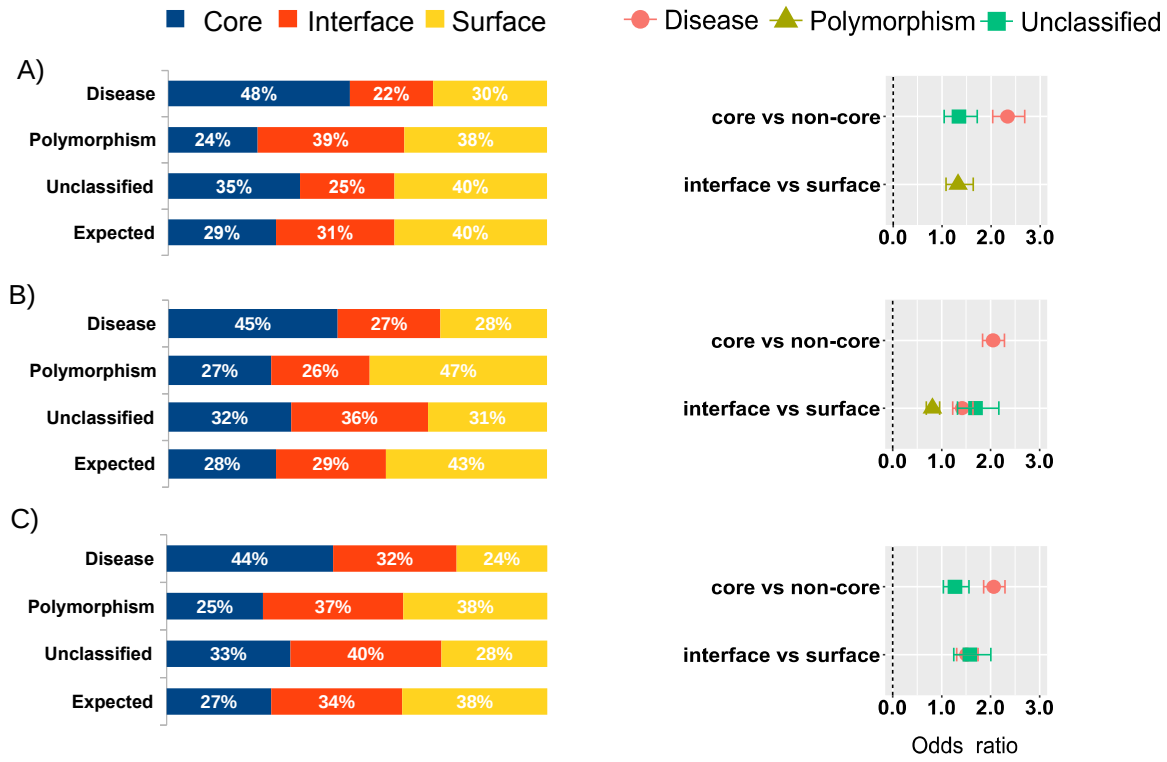



*Figure 16:* ***Distribution of nsSNPs in the protein interaction networks of six selected diseases***

Distribution of nsSNPs (detailed for disease, polymorphism and unclassified) in the protein interaction networks from the six selected diseases, as classified in core, interface and surface non-interface, with expected distributions were calculated from residue composition, and O/E ratios for the different residue locations and types of nsSNPs, based on (A) structural data; (B) modelled interactions; and (C) combined structural data and modelled interactions.

Thus, it remains to be seen whether having more structural data on the protein interactions for these six diseases analyzed here could improve the structural and functional characterization of known disease-related nsSNPs. The following section will explore computational ways to extend the structural characterization of protein interaction networks.

## 4.11 Prediction of interface residues by docking

The main goal of this part is to explore computational ways of characterizing pathological mutations possibly involved in protein-protein interactions for which there is no available structural

data. We previously found that energy-based protein docking can be efficiently applied to identify interface and hot-spot residues in protein-protein complexes (Grosdidier and Fernández-Recio 2008). This approach was implemented in the pyDockNIP module within our docking protocol pyDock (Cheng, Blundell, and Fernandez-Recio 2007). We have evaluated the predictive capabilities of this method at different NIP cutoff values, on the protein-protein docking BM4.0, and the results **(Figure 17)** confirm that this method can predict interface residues with high precision (65-70%), but very low sensitivity (less than 10%). This sensitivity level is too low for its applicability at large protein interaction networks, given that the majority of pathological mutations involved in protein interfaces would not be detected. In order to improve its applicability, we have extended the predicted interface patches by including residues in the vicinity of the originally predicted ones (see Chapter Methods section 3.10). This strategy showed a better trade-off between precision and sensitivity, with improved sensitivity up to 28%, at the expense of precision (**Figure 17**).

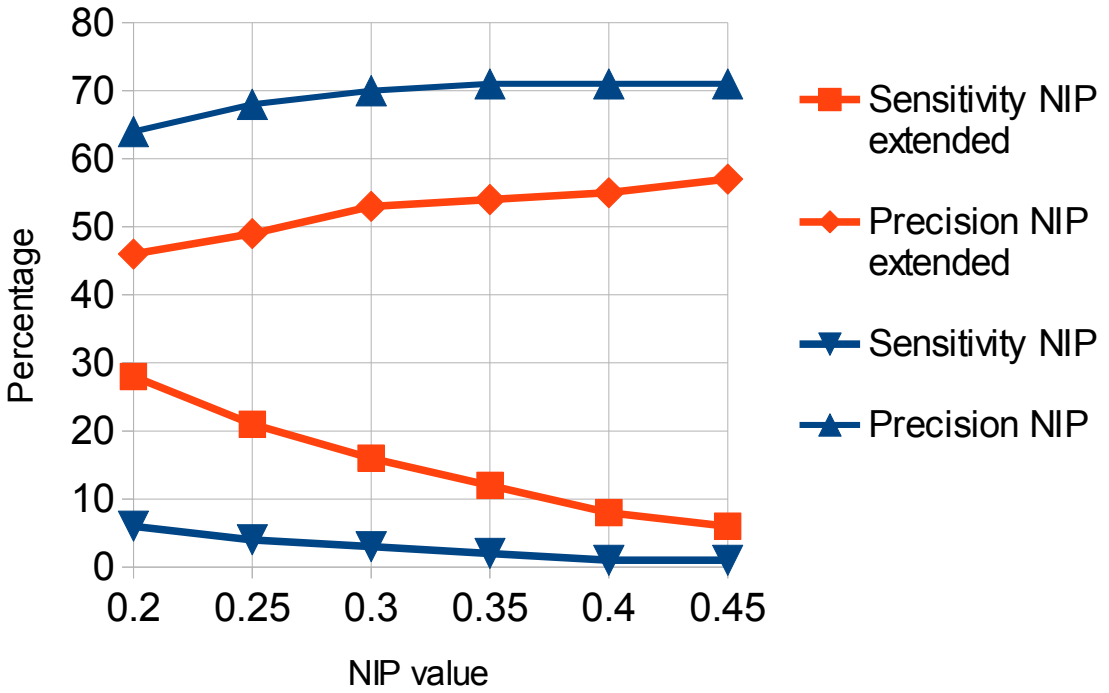



*Figure 17*: ***Prediction of the interface in the BM 4.0 using the extended interface with pyDockNIP and pyDockNIP extended .***

Sensitivity and precision of interface residue predictions based on pyDockNIP (alone or by adding neighbor residues) for proteins in the protein-protein docking BM4.0, according to NIP cutoff value.

As an additional test, we applied this procedure to the structural interaction networks of six selected diseases, as above mentioned, containing 449 protein-protein interactions for which the complex structure is available or can be modelled based on a homologous template, and 353 protein with available structure or model. Some of the proteins in this dataset had more than one binding partner, so we considered as interface residues those that are involved in any of the possible interactions. As a consequence, 44% of the surface protein residues were located at a protein-protein interface (**Table 4**). Then computational docking was run on the separated complex components of the 449 protein-protein complexes, being them either x-ray structures or homology-based models, and the predictions were compared to the real interface residues. The predictions yielded a precision of 64%, with a sensitivity of 50% (**Figure 18**). This improvement in the predictive success rates with respect to the results in the protein-protein docking benchmark might be due to the fact that many of the proteins in the disease-associated interaction networks showed several binding partners, and thus the proportion of surface residues that are at the interface in that set (44%) was larger than in the docking benchmark (23%).

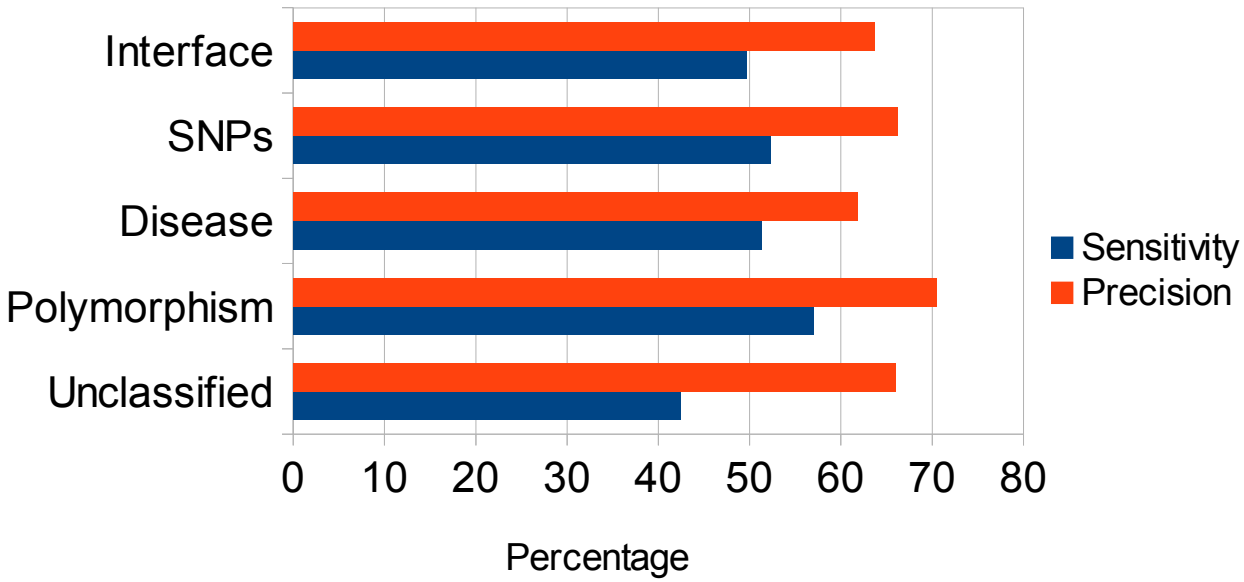



*Figure 18**: Success rates in the mapping of nsSNPs with the predicted extended interface.***

The docking-based extended interface predictions were applied to the structural interaction networks from the six selected diseases. Precision and sensitivity are shown for interface residue predictions and interface nsSNPs. The latter are also detailed for interface disease-related, polymorphism and unclassified nsSNPs.

## 4.12 Docking-based interface prediction can help to improve nsSNP characterization

We tested the docking-based extended interface predictions for the identification of interface nsSNPs in the disease-associated interaction networks, and the predictive success rates were similar to those of the interface predictions (**Figure 18**). We then evaluated how many of the disease-related nsSNPs that are known to be at protein-protein interfaces can be detected by the above mentioned extended interface prediction based on docking calculations. Thus, for all 832 disease-related nsSNPs in our structural interaction network dataset, we applied our docking-based method to predict whether they were located at interfaces. When compared with the 183 disease-related nsSNPs that were actually located at interfaces in our structural dataset, the predictions showed very similar numbers in precision (62%), and sensitivity (51%) to those for the interface prediction (**Figure 18)**. When applied to other types of nsSNPs, the prediction success rates were also similar, except for the "unclassified" nsSNPs, for which sensitivity is slightly lower (**Figure 18**). In general, the above results show that docking-based predictions can identify with reasonable precision when a given nsSNP is located at a protein-protein interface, independently on whether such nsSNP is associated to a disease or no. This provides a valuable resource to characterize nsSNPs in cases with no structural information on the potential protein-protein interactions.

***Table 4: Detailed analysis of location of nsSNPs based on complex structures and modelled interactions***

| | Structures | Models | Combined |
|---|---|---|---|
| **Total analysed proteins** | 353 | 583 | 603 |
| **Tota proteins with At least 1 SNP** | 258 | 411 | 424 |
| **SNPS in interface:** | 449 | 736 | 975 |
| **SNPS in core** | 619 | 970 | 1003 |
| **SNPS in surface:** | 556 | 909 | 808 |
| | RESIDUES | | |
| **Total residues** | 76168 | 189629 | 199846 |
| **Residues in core** | 21710 | 53849 | 54936 |
| **Residues in surface** | 30679 | 80749 | 76142 |
| **Residues at interface** | 23779 | 55031 | 68768 |
| **Total residues hotspot** | 5918 | 11839 | 16449 |
| **Total residues hotspot at interface** | 3673 | 11839 | 14459 |
| | SNPS | | |
| **ALL_SNPS** | 1624 | 2615 | 2786 |
| **Disease** | 832 | 1363 | 1438 |
| **Polymorphism** | 499 | 851 | 899 |
| **Unclassified** | 293 | 401 | 449 |
| | SURFACE | | |
| **Disease** | 250 | 384 | 343 |
| **Polimorphism** | 188 | 399 | 340 |
| **Unclassified** | 118 | 126 | 125 |
| | CORE | | |
| **Disease** | 399 | 609 | 629 |
| **Polimorphism** | 118 | 231 | 228 |
| **Unclassified** | 102 | 130 | 146 |
| | INTERFACE | | |
| **Disease** | 183 | 370 | 466 |
| **Polimorphism** | 193 | 221 | 331 |
| **Unclassified** | 73 | 145 | 178 |
| | HOTSPOTS_INTERFACE | | |
| **Disease** | 33 | 74 | 109 |
| **Polimorphism** | 46 | 35 | 76 |
| **Unclassified** | 17 | 44 | 61 |

## 4.13 Identification of interface nsSNPs in complexes with no available structure

The above described protein interaction networks for the six selected diseases contained 1,485 interactions for which there is no available structure. They involved as many as 3,323 nsSNPs that could not be structurally mapped in such interactions. Some of these nsSNPs might have been considered in the previous analysis of the structural interaction network dataset, simply because they were involved in other complexes with available structure, but they still lacked information for all the other interactions with no available structure. In 1,367 of these interactions, the interacting subunits had available structure or could be easily modelled by homology, which made them suitable for docking calculations. In total, there were 583 proteins with structure or easily modelled by homology (**Table 4**). We ran docking simulations on these interactions to predict interface residues, and then used this information to identify nsSNPs located at protein-protein interfaces. Some of the interacting proteins have different PDB structures corresponding to different parts of the protein, in which case we used all of these structures independently in docking. For instance, in the interaction between the oncogene RAF1 and the heat shock protein HSP90AA1, there are five different PDB structures associated to RAF1, covering different zones of the protein, and two different PDB structures associated to HSP90AA1. Such discontinuous structural coverage for these proteins makes that the modeling of this interaction alone needs 10 independent docking simulations. As a consequence, we run a total of 9,204 docking simulations, and as many of 2,615 nsSNPs could be characterized in 1,367 modelled protein-protein complexes. Within these nsSNPs, we found 1,363 disease-related, 851 polymorphisms, and 401 unclassified. Among the docking-based characterized disease nsSNPs, 45% were buried, 27% were located at a protein-protein interface, and 28% at a non-interacting region (**Figure 16B**). According to the residue composition of the studied proteins, the values expected by chance for buried, interface and non-interface residues are 28%, 29%, and 43%, respectively. Thus, the O/E ratios for buried, interface and non-interface disease nsSNPs are 1.57, 0.94 and 0.66, respectively. These numbers are virtually the same as those found in previous studies on larger interaction sets (1.58, 0.96, and 0.71, respectively (David and Sternberg 2015) ). Based on the modelled interactions, disease nsSNPs have clear preference for being at protein-protein interfaces as compared with non-interacting surfaces (OR 1.42), also in line with previous studies (OR 1.35 (David and Sternberg 2015)). This shows that modeling interaction networks by docking has the capability of extending the characterization of nsSNPs in cases with no available structural data.

## 4.14 Integrated experimental and computational characterization of protein interaction networks

Then, we combined the results of the structural dataset with the modelled interactions for the protein interaction networks of the six selected diseases. In this way, we had structural or modelled data for a total of 2,786 nsSNPs in 2,043 protein-protein interactions. They contained 1,438 disease-related, 899 polymorphisms and 449 unclassified nsSNPs. Among the characterized disease-related nsSNPs, 44% were buried, 32% were located at interfaces, and 24% at non-interacting regions (**Figure 16C**). According to the residue composition of the structurally characterized and modelled proteins, the values expected by chance for buried, interface and non-interface residues are 27%, 34%, and 38%, respectively. Thus, the O/E ratios for buried, interface and non-interface disease nsSNPs are 1.59, 0.94 and 0.63, respectively (similar to previous studies David and Sternberg 2015)). This shows an even clearer preference of the disease nsSNPs for being at interfaces rather than at non-interacting regions (OR 1.51). This clearly shows that the combination of experimental and computational information can help to improve the structural characterization of protein interaction networks and the identification of nsSNPs involved in interactions.

Interestingly, the disease-related nsSNPs that are estimated to be interacting hot-spots according to the docking-based predictions (interface residues with NIP > 0.2) show an O/E ratio of 1.05, and a clear preference over the non-interacting regions (OR 1.68), similar to that previously reported for interface core disease nsSNPs vs. non-interacting regions (OR 1.72 (David and Sternberg 2015)).

For the interaction networks of the six selected diseases, on top of the 183 disease-related nsSNPs that could be structurally mapped at protein-protein interfaces, we found 283 additional disease-related nsSNPs that were predicted to be at an interface based on the docking models, yielding a total of 109 interface disease-related nsSNPs that were also characterized as hot-spots, and which are likely to have a significant edgetic effect.

## 4.15 Docking-based characterization of pathological mutations in the RAS/MAPK pathway

We used our interface prediction method to extend the characterization of nsSNPs in other protein interaction networks. A recent comprehensive study on pathological mutations involved in cancer and RASopathies in proteins of the RAS/MAPK pathway showed that around 20% of the structurally-mapped pathological mutations were predicted to have a direct effect on protein-protein or domain-domain interfaces(Kiel and Serrano 2014). However, for over 30% of the mutations that could be mapped at a protein structure, they could not find any structural or energetic reason that might explain their pathological character. The majority of these mutations were located at the protein surface, and the authors proposed that they might be involved in protein interactions for which there is no sufficient structural data. Some of the mutations could be located at a known protein-protein interface but perhaps do not have any impact on the binding affinity (Teng et al. 2009), while they could actually affect other protein-protein interactions with no available structural data (Keskin and Nussinov 2007; Martin and Lavery 2012). Therefore, we aimed to complete the interface structural and energetics data of this protein interaction network with our computational approach, to explore whether this can help characterizing some of these "unexplained" mutations. We used the first-degree neighbors to construct the network for the 15 proteins analyzed in the mentioned study (Kiel and Serrano 2014).

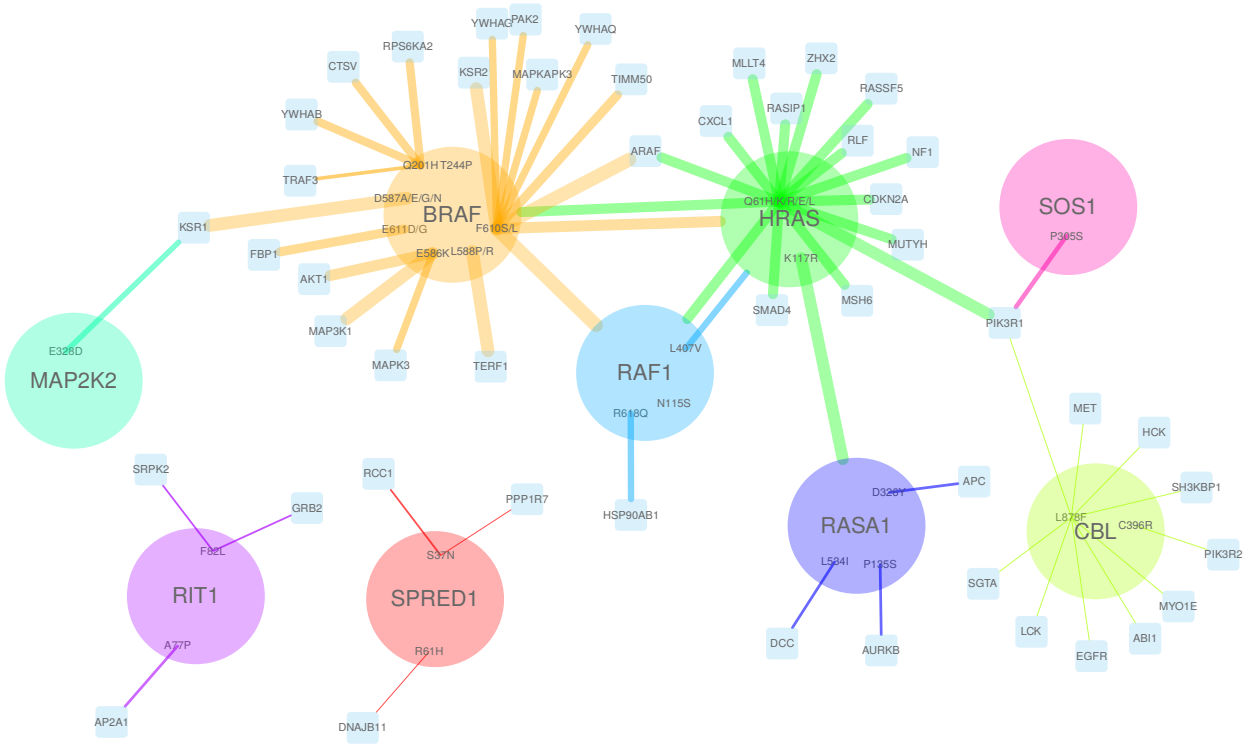



*Figure 19:* **Structurally unexplained mutations of the RAS/MAPK pathway that are predicted to be involved at protein-protein interfaces**

The mutations shown here were not previously characterized due to the lack of structural data, but have been predicted here to be involved in protein interactions as hot-spots, based on docking calculations (see Methods). Main proteins of the RAS/MAPK pathway are represented as circles, and the additional proteins in the first interaction layer are shown as cyan squares. The interactions affected for each mutation in the main proteins are shown as connecting links. The thickness of the edge line is related to the number of pathways in which a given interaction is involved

The complete interaction network involved a total of 236 proteins, 234 of them with available structure, and 482 protein-protein interactions (300 of them without structural information). We performed 1,893 docking calculations on those protein interactions with no available structure, in order to identify the interface and hot-spot residues. From the 208 nsSNPs that were unexplained in the mentioned study (David and Sternberg 2015), we found 95 nsSNPs (in 59 residues of 11 proteins) that were predicted to be at a protein-protein interface based on the docking calculations. That is, interface predictions based on docking calculations helped to rationalize almost half of the unexplained mutations. Among them, we found 44 nsSNPs (in 29

residues of 9 proteins) that were predicted to involve a protein-protein hot-spot residue **(Figure 19)**. These nine proteins play a significant role in the Ras pathway, and are found to interact with several other signaling proteins. Cross pathway connectivity among signaling proteins is a network property that is related to the robustness or fragility of cell functions (Martin and Lavery 2012). Therefore, nsSNPs located at protein-protein interfaces in these nine proteins could not only affect the Ras pathway but also other pathways. Figure 20 hows the pathways involving proteins whose interaction is affected by the pathological mutations predicted to be located at a binding hot-spot. We found the most affected pathways are related to the vascular system formation and activation of immune cells. The VEGF, PDGF, FGF and interleukin signaling pathways are closely involved in cell proliferation, differentiation and angiogenesis, all of them highly relevant in the development of cancer.

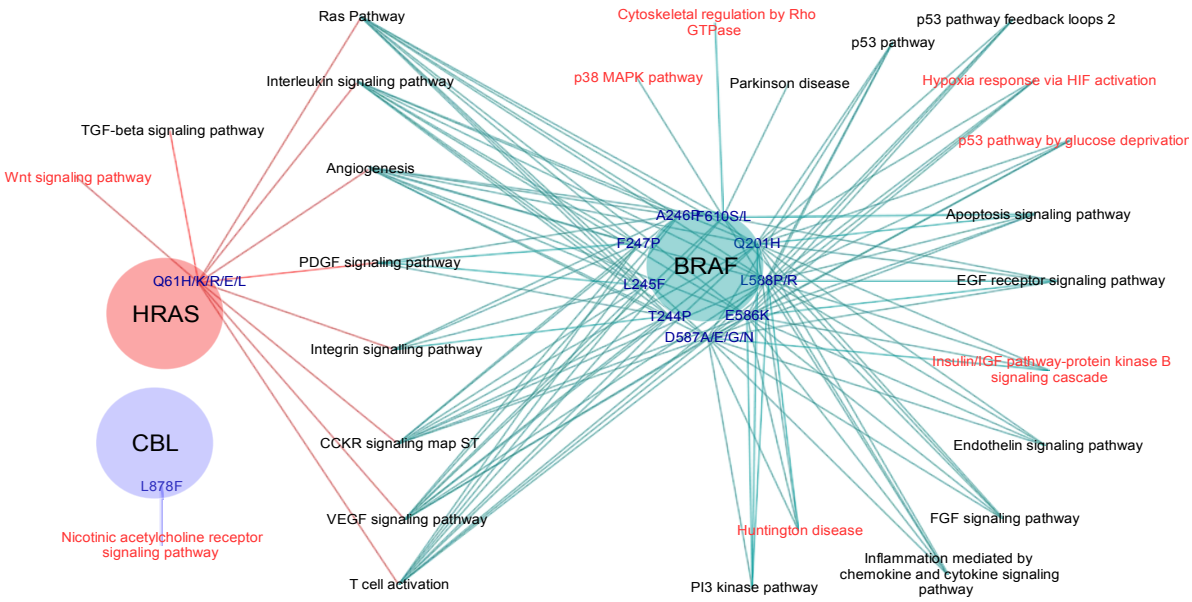



*Figure 20:* ***Pathways affected by pathological mutations in RAS/MAPK proteins predicted to be at binding hot-spots***

Proteins of the RAS/MAPK pathway are shown as colored circles, showing pathological mutations that were not previously characterized due to the lack of structural data, but that have been predicted here to be binding hot-spots for docking partner proteins involved in other pathways (linked to the corresponding mutation).

## 4.16 Interactome and core diseasome analysis with high-throughput docking simulations

In previous sections of this thesis, we have shown how docking-based interface predictions can help to characterize mutations in selected interaction networks that vary in size and complexity. Here, we have extended our protocol to analyze the entire high-confidence human interactome (Rolland etal 2014) and the human core diseasome (Janjić, and Pržulj. 2012). In total, we analyzed 4,254 different proteins involved in a total of 11,925 interactions and 14,551 SNPs from the humsavar data file. In addition to the analysis previously performed in six selected disease networks, we also used the ZDOCK docking protocol to predict the interface residues. From the scoring function analysis in section 4.8 we found that combination of PYDOCK_TOT with ZDOCK gives a high-success rate without the risk of overfitting. One of the reasons for using this additional strategy is that some complementarity to the pyDock docking protocol was previously observed when evaluating

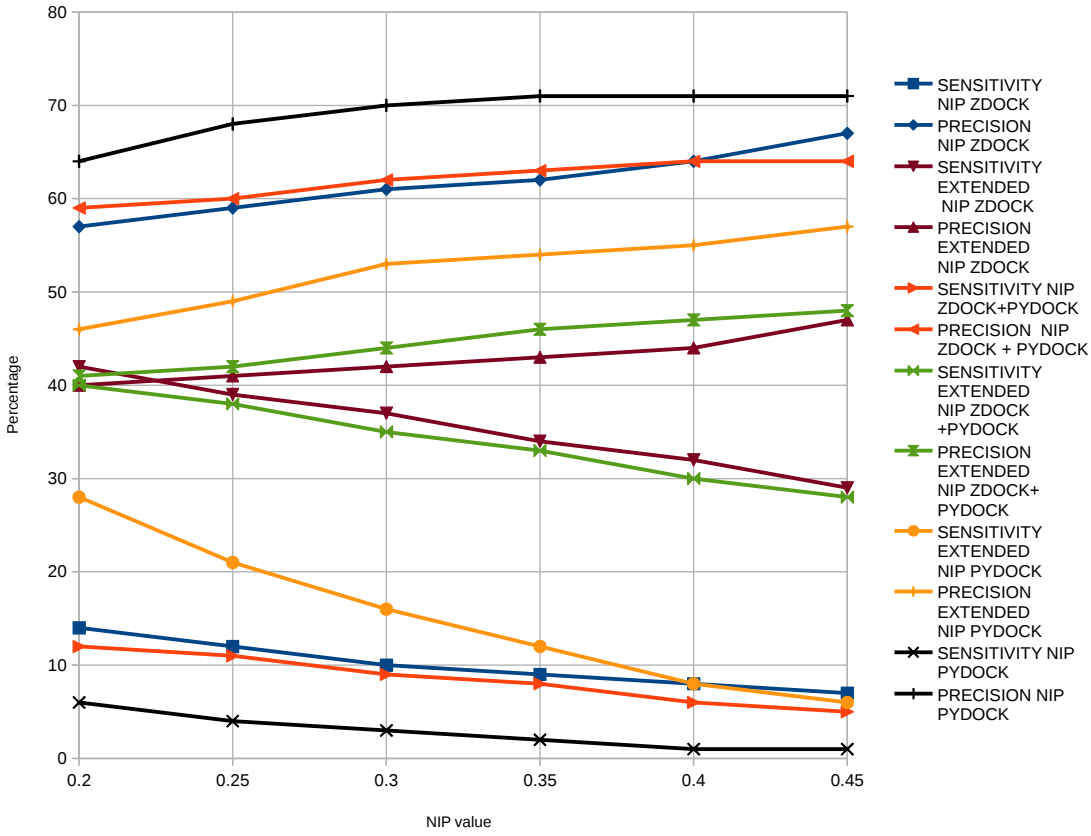



*Figure 21:* ***Sensitivity and precision comparison between ZDOCK and FTDock based extended NIP***

ZDOCK decoys with PYDOCK_TOT (Mosca et al. 2009). We selected the highest NIP value found by any of the docking methods, in order to predict as many hot-spots residues as possible. Taking into account all the docking runs performed in the human interactome and the core diseasome, we performed 36,678 dockings simulations for FTDock and 36,551 for ZDOCK .The difference in the successful number of docking runs comes from the fact that our version of FTDock is optimized to run in parallel processors using shared memory this allows the discretization of the complicated geometry of the protein into the grid, such as large lineal helices, while ZDOCK run in a singe processor limited to the memory available to that processor. After, we evaluated the decoys generated by these two docking method with our scoring functions PYDOCK_TOT. **Figure 21** shows and compares the sensitivity and precision to find residues at the interface in the BM 4.0, using ZDOCK Scoring and ZDOCK evaluated with PYDOCK_TOT and the pyDockNIP from FTDdock. Clearly, the sampling method influence the sensitivity and precision of the hotspot prediction and the NIP extended. In the NIP alone prediction, FTDock has greater precision and ZDOCK has higher sensitivity that drops to the same level as FTDock as cutoff value increases. In general, both types of NIP precision obtained from ZDOCK is in the range of 40-50% while the sensitivity is around 30-40%. This is a relevant difference with FTDock, in which precision increases with the cutoff values. Still, the most beneficial cutoff value for NIP extended generated from ZDOCK is 0.2.

We analyzed the high-confidence human interactome and core diseasome interaction networks, in the same way as we previously analyzed the six selected disease networks(see section 4.15 in this chapter). **Figure 22** shows the sensitivity and precision for the method of the maximum NIP and neighbors to find the three types of SNPs at the different zones, assessed on the interactions with available structure. This test was performed on a set of 2,226 PPIs with crystal structure from the PDB or reliable homology-based model. The observed results are different to the results obtained from the six diseases, due the sample size difference and use of other docking method.

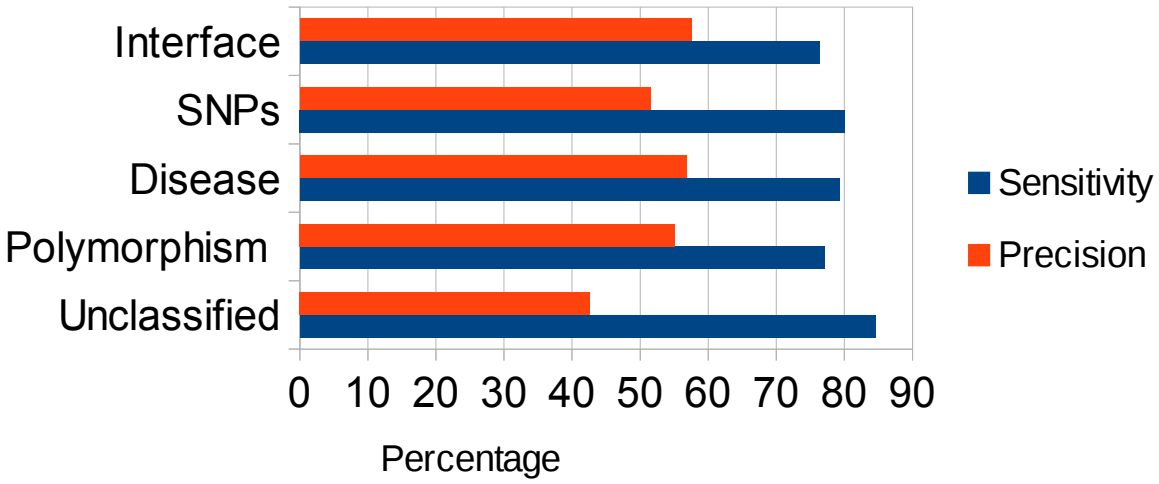



***Figure 22: Sensitivity and precision of the NIP extended methods in all the complexed protein structures***

This interface prediction strategy using the maximum NIP and extending it to the neighbors shows a sensitivity and precision above 75% and 50%, respectively, for the interface residue and SNPs predictions (except for unclassified SNPs). Thus, the interface residue predictions reach a sensitivity of 76% and a precision of 58%, while the interface disease-associated SNPs show a sensitivity of 79% and precision of 57%.

Next, we analyzed the distribution of the nsSNPs in these interaction networks, including the interactions for which there is no available structure. **Table 5** shows details of the analysis carried out.

***Table 5: Detailed analysis of location of nsSNPs based on complex structures and modelled interactions for the human interactome and core diseasome***

| | Structures | Models | Combined |
|---|---|---|---|
| Total number of proteins | 2039 | 4022 | 4254 |
| Proteins with known nsSNPs | 1293 | 2549 | 2712 |
| | | | |
| Interface nsSNPs | 2500 | 4309 | 6261 |
| Core nsSNPs | 1583 | 3970 | 3935 |
| Surface nsSNPs | 2706 | 3977 | 4355 |
| | | | |
| Total residues | 409219 | 1162790 | 1231941 |
| Core residues | 78571 | 330447 | 306001 |
| Surface residues | 173014 | 421402 | 423973 |
| Interface residues | 157634 | 410941 | 501967 |
| | nsSNPS | | |
| Disease | 3259 | 6145 | 7326 |
| Polymorphism | 1640 | 4153 | 4461 |
| Unclassified | 1890 | 1958 | 2764 |
| Total number of nsSNPs | 6789 | 12256 | 14551 |
| | SURFACE | | |
| Disease | 1054 | 1583 | 1881 |
| Polymorphism | 793 | 1926 | 1833 |
| Unclassified | 859 | 468 | 641 |
| | CORE | | |
| Disease | 992 | 2546 | 2573 |
| Polymorphism | 225 | 887 | 804 |
| Unclassified | 366 | 537 | 558 |
| | INTERFACE | | |
| Disease | 1213 | 2016 | 2872 |
| Polymorphism | 622 | 1340 | 1824 |
| Unclassified | 665 | 953 | 1565 |
| | INTERFACE HOTSPOT nsSNPs | | |
| Disease | 279 | 716 | 949 |
| Polymorphism | 156 | 388 | 547 |
| Unclassified | 205 | 333 | 584 |

**Figure 23C** shows the location of SNPs in the interactome and core disease networks, based on the available structural data and on the docking-based interface predictions. **Figure 23A** shows the same analysis using structural data for only the interactions with available structure. **Figure 23B** shows the analysis for the interactions with no available structure, with interface predictions based on the docking models. We can observe the expected strong preference of disease-associated SNPs for core region vs. non-core regions. It is known that many pathological mutations cause the disease by affecting the folding and stability of a protein. Although to a lesser extent, we also find a preference of disease-associated SNPs for the interface zone vs. non-interacting surface. The

polymorphism nsSNPs have a distribution closer to the expected one, showing no preference for core or interface regions.. Interestingly, the unclassified SNPs show an important preference for the interface vs. non-interface surface while they do not show preference for core. These could be disease-related SNPs that do not have a strong effect on the structure of the protein, but could have a more subtle effect by affecting specific interactions, which together with the limited number of patients from which they are derived, make that they are difficult to unequivocally associate to a pathological condition. . Interestingly, when the predictions were assessed on the interactions with available structure, the unclassified SNPs showed the lowest precision. This high rate of false positives could indicate that these SNPs are affecting interactions for which there is no available structure, and that is another reason why they have not been reliably associated to a disease yet. Of course, we should not disregard that the false positives found in the prediction benchmark might just be an artifact of the prediction of this type of SNPs, so these results should be taken with caution.

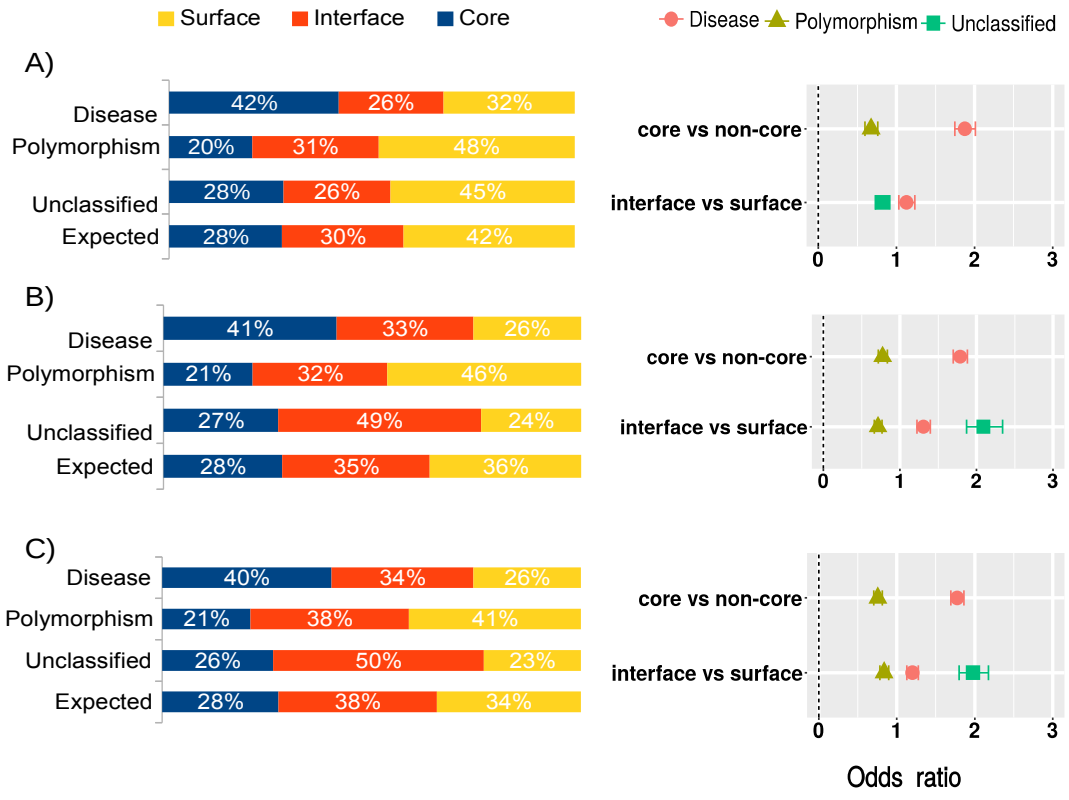



*Figure 23:* ***Distribution and odds ratio for the docking-based interface predictions in interactome and core diseasome***

See details in Figure 13

The network generated for the human interactome with the interactions affected by SNPs based on the structures and docking models is difficult to visualize in a printed figure due the large number of interactions in the interactome and core diseasome networks. This network has 585 nodes, corresponding to 3,346 edges representing the PPIs. About 1,284 interactions are affected by at least one disease-associated nsSNP, and 1,349 are affected by unclassified nsSNPs. We found 21 pathways can be affected by this mutations. **Figure 24** shows a simplified network derived from our analysis, showing only the interactions affected by each type of nsSNPs.

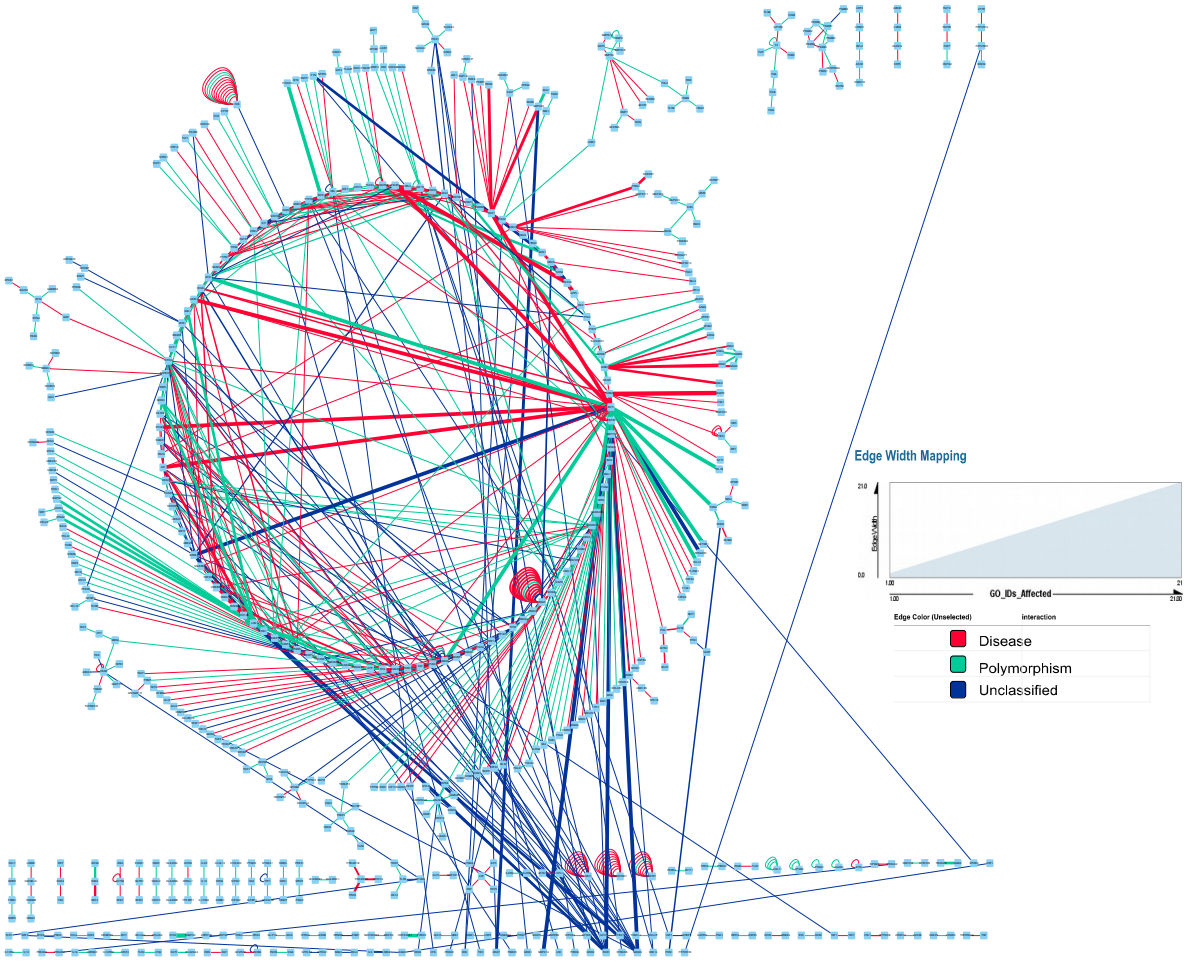



*Figure 24**: Simplified network of the human interactome network affected by nsSNPs at the interaction.***

A representation of the interactome organized in a circular layout. Nodes in blue squares represent protein where nsSNPs are predicted to be at the interface of a PPIs.

To complement this result we analyzed the molecular function associated for all the proteins involved in an interaction that are affected by diseases-asociated nsSNPs according the GO classification. As expected, the analysis show an over-representation of the binding of proteins in different context such enzyme binding, nucleotide binding, cytoskeletal protein binding. This means

that a broad spectrum of this functions are altered by the nsSNPs that cause a disease. **Figure 25** show the most representative cluster of the over-representation analysis.

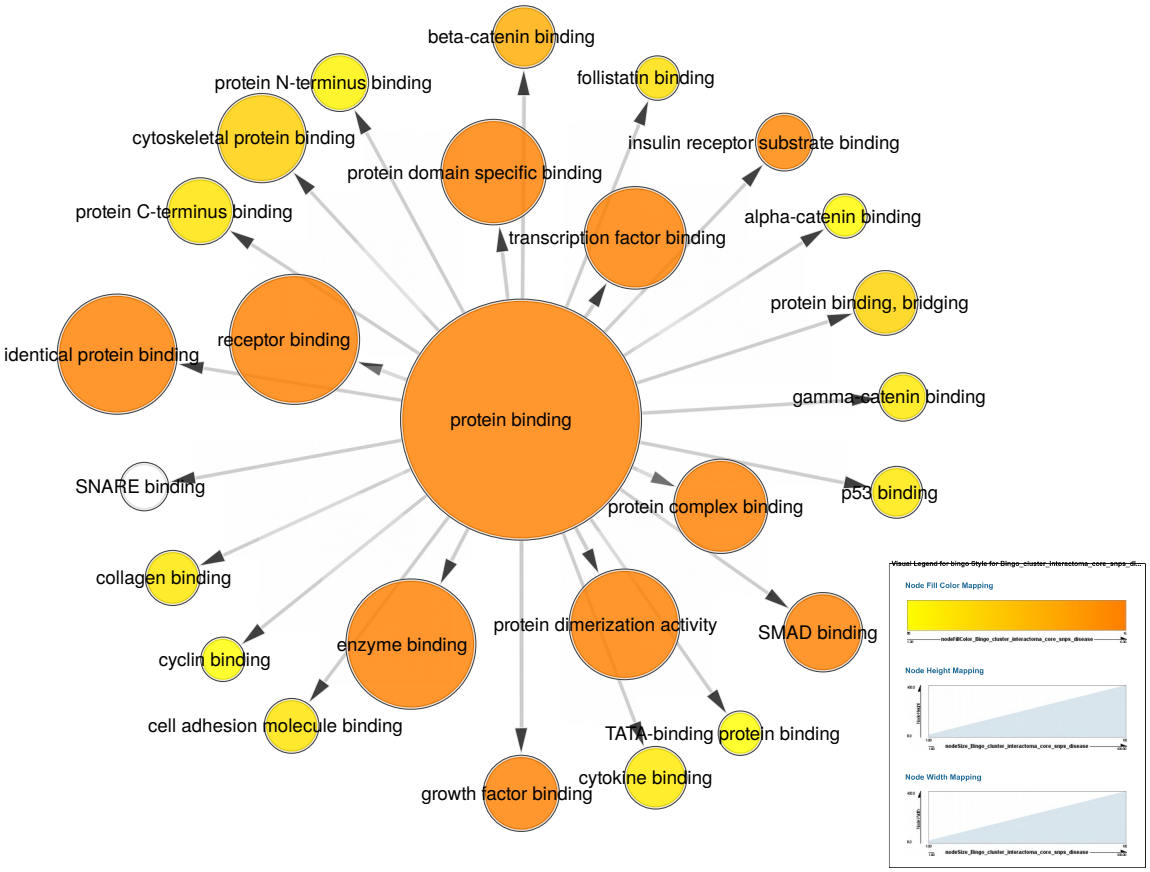



*Figure 25:* ***Overrepresentation analysis of GO molecular functions altered by the disease nsSNPs at the interface.***

Size the node indicate the enrichment of the GO molecular term associated to the name of the proteins in the network. At the center, the GO term of Protein binding is the parent term of all the other GO terms, arrows indicate this hierarchic array . Color indicate statistically significance associated to the GO term. The darker the color the higher their statistical significance.

Going further with the analysis, we clustered the generated network using the docking models and structures according to the edge betweeness metric. This metric is an indicator of the importance of the interaction given the topology of the network. **Figure 26** show the cluster with the top scoring clustering in the network.

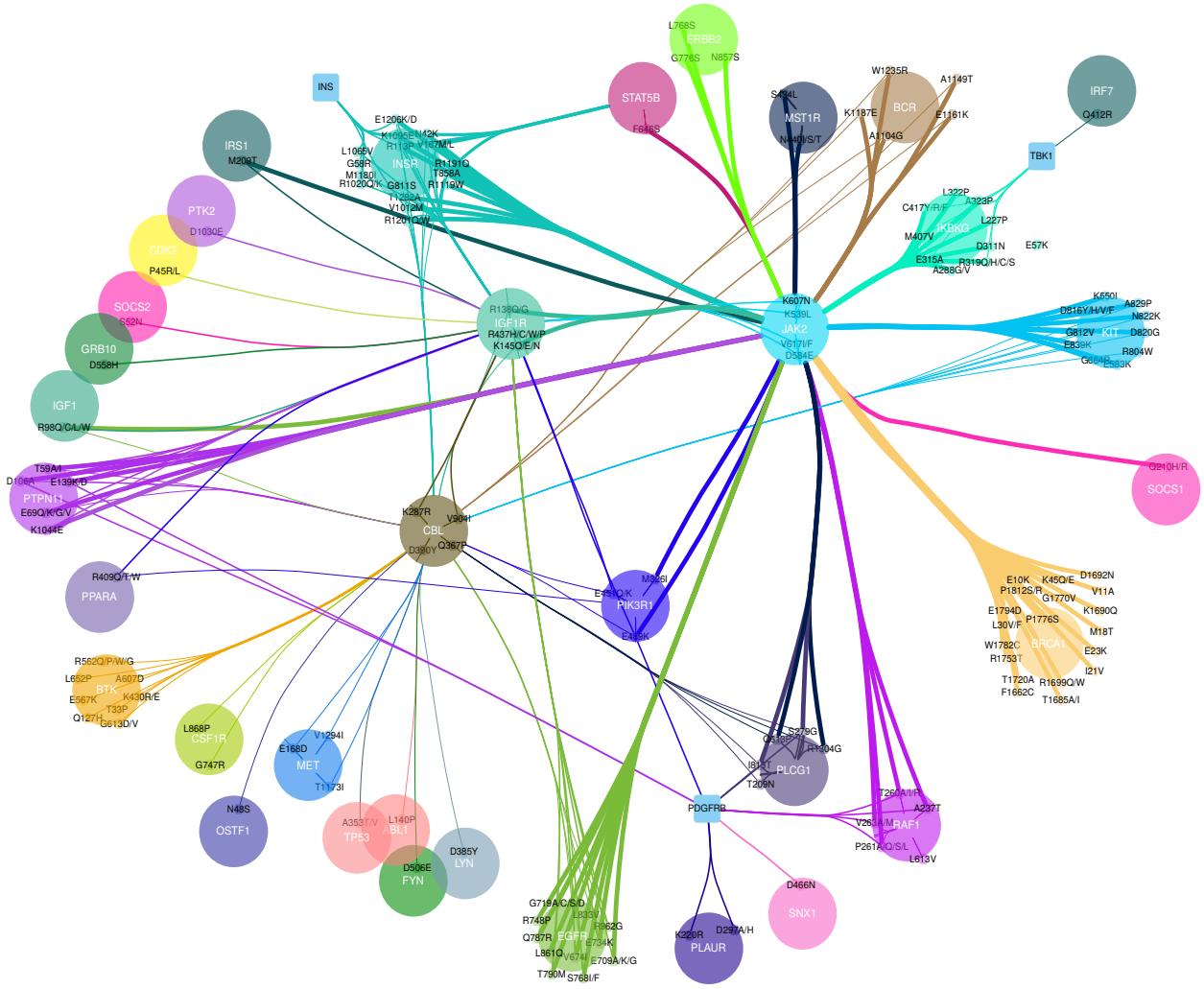



*Figure 26:* ***Top scoring cluster of PPIs according the to edge betweeness metric***

Edge colors correspond to the node (protein) of origin. Edges are bundled together to facilitate the visual representation of the modules formed by the PPIs.

The topology of the network shows different modules produced by the interaction between proteins with nsSNPs at the interface zone. The network view with bundle edges helps to analyze the burden caused by the altered interactions. For instance, the protein PDGFRB does no harbor any nsSNP at the interface, but it is surrounded by different proteins with at least 1 disease-associated nsSNP. All the interaction partners are increasing the burden of their nsSNPs in this particular protein. Also we observed some modularity in the network. There is a complex module established by the interactions between JAK2, RAF1, KIT, CBL, INSR, IGF1R, BCR, STATB5, PLCG1, EGFR, PPARA, and PTPN11. Three of these 12 proteins, JAK2, CBL, and IGF1R, are surrounded by a large number of neighbors, which means that the burden of the nsSNPs is larger in these three

proteins than in other parts of the interactome.

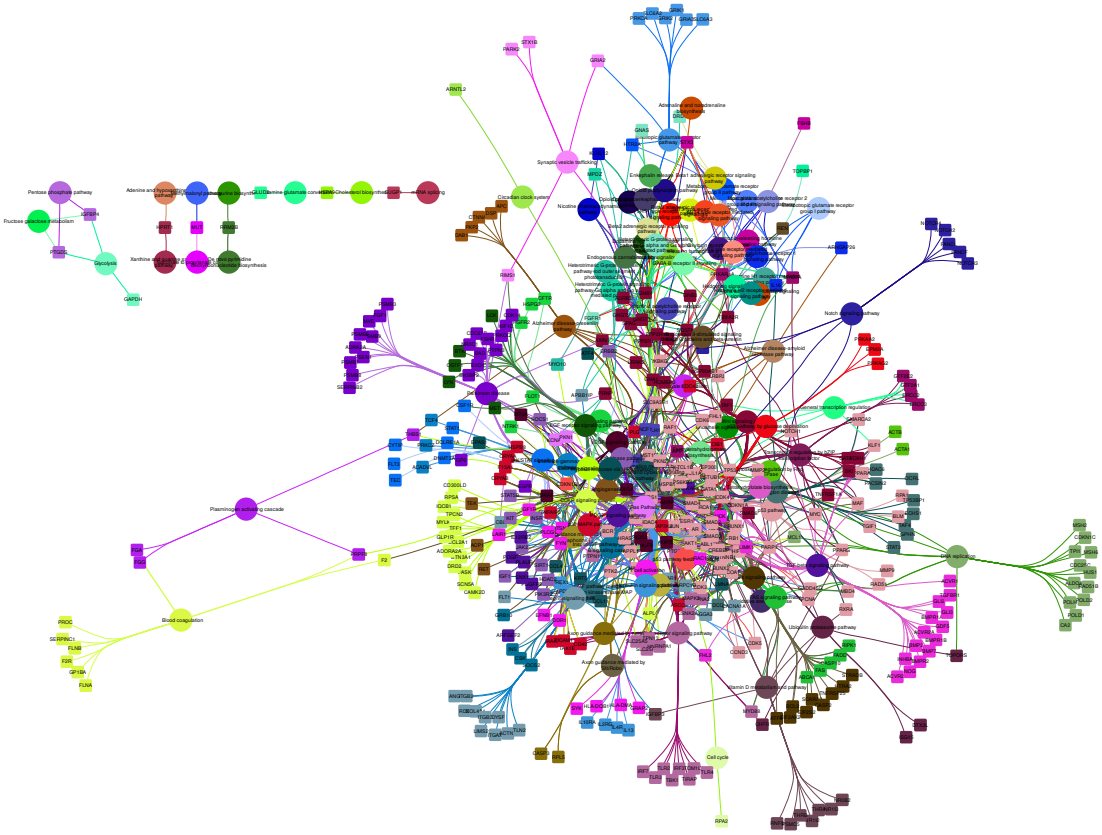


*Figure 27:* ***Cellular pathways affected by nsSNPS at the interface of the proteins in the human interactome***

The interaction network all pathways (circles nodes) predicted to be affected by disease nsSNPs, because they involve proteins (square nodes) whose interaction is altered by disease-associated nsSNPs according to the interface predictions and available structural data. The color represents the pathway in which each protein is involved. Edges are bundled to maximize the visualization of the clusters formed by simple association between proteins and pathways.

As in the previous study of the pathological mutations in the RASopathy networks based on the docking calculations, we focused the analysis to the cellular pathways that are possibly affected by the nsSNPs that are found at the interface of the proteins. **Figure 27** shows all the possible pathways that are affected by the existence of a nsSNP at the interface of a protein. **Figure 28** shows the biggest cluster of pathways according to the edge betweeness measurement. In this cluster the central and most involve pathway is CCKR signaling map ST. Most of the interaction partners of the displayed protein seem to have a role in this pathway and probably affecting another

14 pathways out of the 17 total pathways. A characteristic of this network is also that many of the proteins in this cluster have a maximum of two interaction partners. This indicate to some degree the burden of the nsSNP on the protein. For example, disease nsSNPs in KAT5 at the interface seem to affect 5 different pathways.

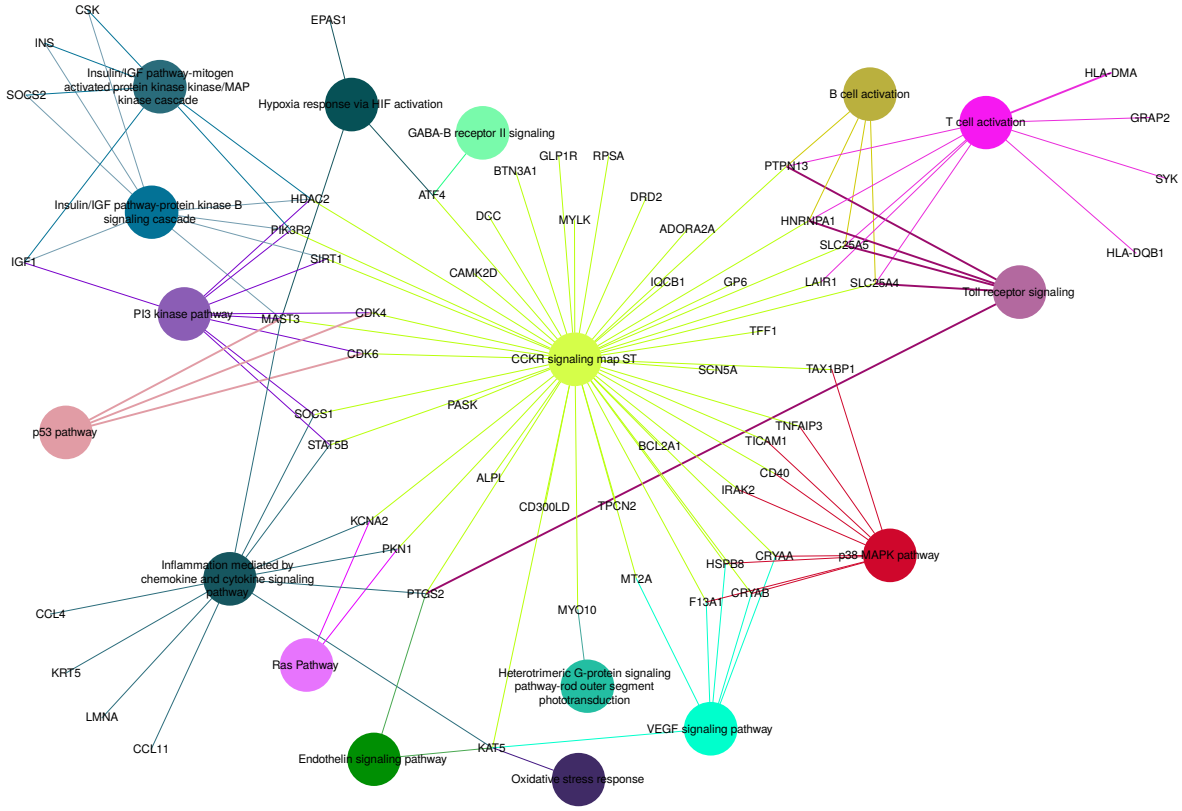



*Figure 28:* ***Top scoring cluster with the highest edge betweeness from the pathway analysis***

17 different pathways are predicted to be the most affected using automatic clustering by edgebetweeness. The network show in circular nodes the pathways and the name of protein involved in altered PPIs appears at the intersection of the edges coming from the different pathways

This network analysis helps to find other proteins that are relevant in terms of increasing the burden of a disease nsSNPs on cellular pathways rather than in specific partners. In addition, we can observe some degree of connection between similar cellular functions, such as in the cases of the Toll signaling pathway, inflammatory pathways or T/B Cell activation.

“The first rule is to keep an untroubled spirit.

The second is to look things in the face and know them for what they are.”

— Marcus Aurelius

# Chapter 5 Discussion

## 5.1 Insights from post-docking analysis in rigid body sampling

Proteins are the physical embodiment of the message contained in the genes and they often work through protein-protein interactions that are critical in all cellular processes. The study of the structural and energetics requirements for these protein interactions is fundamental to understand better biological processes at molecular level and to develop new therapeutic applications. Computational docking can help to overcome the technical limitations in experimental structure determination. A key component of a docking algorithm is an efficient scoring function capable of identifying the correct docking orientations. Two factors have limited the development of more accurate scoring functions: the availability of source data for training and testing, and the challenge of describing the conformational flexibility of the interacting proteins. The first limitation factor is related the lack of enough protein complexes with a high-quality structure. Provided there were high-quality crystal structures for all types of interactions, then the optimization of the different energy terms in a scoring function would be much more efficient. The second limitation factor remains a major challenge in protein-protein docking, since the efficient description of conformational changes upon binding would require side-chain and backbone refinement during docking, and on-the-fly adjustment of the energy terms. Our post-docking processing of the docking decoys that have been generated from different programs has proven to be a successful method to characterize protein complexes and to explore the differences in sampling. We found several scoring functions that provided better predictions than the inbuilt scoring functions in each program. Moreover, the simple combination of such scoring functions enhanced the detection of near-native solutions. Usually, the best performing pair combination of functions mixed different levels of resolution, e.g. coarse-grain and atomistic. Among the coarse-grain scoring functions, CP_HLPL showed great adaptability to different sampling methods, with no indication of overtraining. In addition, this function performed successfully on the most flexible cases from both benchmarks, but it did not combine well in a pairwise manner with another scoring functions. Combining two scoring functions with different resolutions has been a discussed idea in the docking community. Combining a coarse-grain function with an atomic one would help to pick up different signals

arising from the two different resolution levels. Indeed, we found that many pairs of scoring functions have an acceptable success rate, consistently with the above discussed ideas. Interestingly, some atomistic scoring functions can also combine with other scoring functions of the same resolution level, like PYDOCK_TOT. We pushed further the combinations with the FFT methods, by mixing the best pair of scoring functions for each docking method. The combination of the different pairs for each of the three different methods revealed a slight set of combinations that could be useful for re-scoring the generated decoys. The analysis of cardinalities, such as the symmetric difference metrics, helped to identify that many of the combinations of the three methods had significant overlapping. Nevertheless, a few combinations with the three different methods showed a significant improvement in the predictive success rates, as compared to the simple methods. On the down side, we faced the limitation of not having another external set to test them to reduce the possibility of overtraining.

## 5.2 A single scoring function does not provide an effective description of protein complex formation

A sophisticated framework for integrating a variety of scoring functions is required in order to take advantage of the different signals that all available scoring functions might provide. Determining the relevance of a scoring function for correctly ranking docking decoys is a task that requires several regression and adjustments in a multivariate model. The previous analysis clearly showed us the direct interplay between different scoring functions, and our approach was very direct and simple, but it posed a high risk of overtraining in the BM 5.0. On the other side, although the docking community provided useful benchmarks to assess the success rate of the scoring functions, we faced two drawbacks to using them for the purpose of evaluating the re-ranking power of a variety of scoring functions. The first drawback was the size. For instance, although the scorer set benchmark provided a variety of models from different groups, it was composed of a small number of cases. The second drawback was that it cannot be completely disregarded that any of the scoring functions had been trained on some of the previously available benchmarks. Fortunately, the recently released BM 5.0 update was a good external set that gave us the opportunity to train models and assess their performance, thus avoiding the risk of overtraining the models. By using the BM 4.0 as training setfor the several available scoring functions, together with the quality rank of

each of the thousand decoys generated for the docking programs, and a stepwise feature selection process, we successfully trained hundreds of SVM models. The actual innovation was a final step that used a voting system to obtain a consensus ranking from the best performing models. This combination pushed the near-native poses from each docking method to the top of the list, which improved even further each of the resulting rankings. The success rates of the FFT-based methods were comparable with the success rates of the best general docking methods. Actually, the only method that performed better than the FFT-base methods was SwarmDock when the SVM and voting system was applied to its ranking. Using the enhanced SwarmDock as a point of reference, we noted that the sampling method had a significant influence on the successful identification of a near-native solution. Still, SwarmDock remains computationally expensive and prohibitive for high-throughput use. In BM 4.0 the re-ranked docking procedures beat almost all other methods, except for default Swarmdock, which uses a different sampling strategy. Not all docking methods have the success rate published for the BM 5.0, but the re-ranked FFT-based methods were superior to HADDOCK, one of the best protein-protein docking methods in CAPRI, and were very close to re-ranked SwarmDock. This difference shows again the influence of the sampling on the scoring functions, as well as the advantage of integrating many weighted scoring functions. With this result, we show how protein-protein docking can be further improved, with the use of different biophysical descriptors previously gathered in our group and widespread techniques borrowed from the computer sciences. It remains to be seen if computational strategies such as a deep neural networks can improve the results in the same way as we did. It would be also important to tune the different sets of descriptors according to the difficulty of each case, due to their flexibility or binding energies. Another improvement would be devising a meta-model able to integrate the best of the different FFT-based docking procedures, since these are computationally inexpensive and can search the whole 3D space. The base of this meta-model would be the observation we made trying to combine the normalized values of the three methods, which resulted in a considerable increase of the success rates on the BM 5.0.

## 5.3 The hard task of linking structural information to phenotypes

Structural characterization of nsSNPs and their involvement in protein-protein interfaces is a starting point to understand complex diseases, for which databases like dSysMap (Mosca et al. 2015) are valuable resources. However, a major problem is the limited structural data available for

protein-protein complexes, and as a consequence, only a fraction of all possible nsSNPs can be accurately located at the interfaces. In this work, we have used docking models to characterize nsSNPs that are likely to be involved in protein-protein interactions. To test this approach, we have selected six complex diseases in which their associated proteins are involved in protein-protein interactions for which there is little structural data.

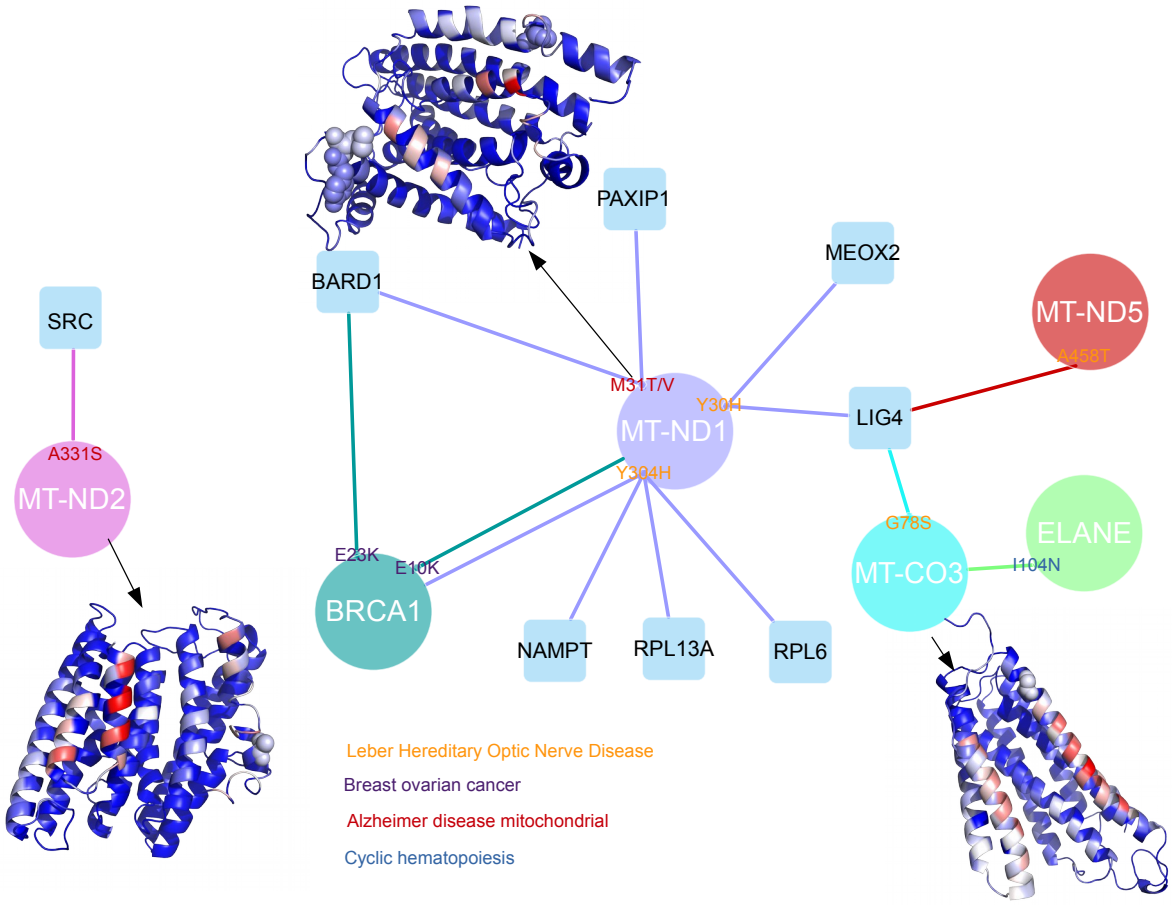



*Figure 29:* ***LHON network with nsSNPS observed with the Structures and pyDock NIP extended predictions.***

Given the limited structural data on the protein interaction network in LHON pathology, docking-based predictions are key to identify disease nsSNPs (CPK representation) located at protein-protein interfaces. Selected protein structures/models are shown, with residues colored according to their NIP value (in red NIP > 0.2; in blue NIP < 0.0).

The first difficulty we encountered in this analysis was the availability of data. The task of finding all the coding protein genes to construct the protein interaction network of a complex disorder is not trivial, as there are different sources of data for nsSNPs (e.g. *humsavar*) and disorder

genes (e.g. OMIM) that are not always fully consistent. Indeed, the gene map file used here from OMIM had 3,438 described phenotypes, while the version of *humsavar* used in this work has 2,727 phenotypes with assigned nsSNPs. This means that there could be protein coding genes associated to a disease phenotype, which do not have any described nsSNP. An example of this was the phenotype MCI (susceptibility to myocardial infarction) [MIM: 608446]. This phenotype is not considered in databases like dSySmap, because all coding protein genes that have been reportedly associated to the disease harbor mutations for other diseases, and thus no nsSNPs can be found associated with this MIM code in *humsavar* (**Table 1**). Therefore, a specific analysis of this phenotype using only the nsSNPs annotated in *humsavar* is not realistic. When we analyzed the interaction network of the proteins associated to this disease, including all nsSNPs associated to any other diseases, we found a strong preference of these nsSNPs to be at an interface rather than in non-interacting regions (OR 1.52, *P*-value < 0.005). The involvement of different nsSNPs causing other diseases in the protein-protein interfaces of this interaction network is indicative of a complex genotype-to-phenotype relationship, which is probably masking the nsSNPs linked to this specific MCI phenotype.

Due to the limited structural data, in phenotypes like the Leber hereditary optic neuropathy (LHON) [MIM:535000], a rare mitochondrial disease, not a single nsSNP related to this disease could be structurally located at a protein-protein interface, since there are no structures for the protein complexes involved in this disease except for the self-interactions. Using our docking-based interface prediction approach, we were able to structurally map 12 of the 21 nsSNPs associated to this disease, and found that 4 of these mutations are predicted to be located at a protein-protein interface; additionally we found nsSNPs associated to other diseases like Alzheimer and Breast-ovarian cancer (**Figure 29**). We found the nsSNPS associated to LHON in three out of six of the protein associated to the disease (MT_CO3, MT-ND1, and MT-ND5). We notice that these proteins are part of the respiratory chain. One of the proteins, MT_CO3 (UniProt P00414), is part of the complex IV assembly of the cytochrome oxidase c, which is the terminal member of the respiratory chain of the mitochondria. The other two affected proteins are components of the NADH-ubiquinone oxidoreductase complex, which is key to the catalytic function of the respiratory chain. We could only analyze part of the chain 1 (MT-ND1, UniProt P03886) and chain 5 (MT-ND5, UniProt P03915). MT-ND1/2 harbor more nsSNPs at the interface that are also linked to Alzheimer's disease (MIM 502500). MT-ND1/5 proteins are involved in the recognition of BCRT

domains, especially MT_ND5. The nsSNPs that we found located in the interface might be very specific for this LHON disorder, probably altering the recognition of such domains. We also found other elements of the respiratory chain affected by nsSNPs at an interface zone, which were described to cause other mitochondrial related disorders. For example, the protein ELANE (P08246), a mitochondrial elastase, is involved in two different diseases, cyclic hematopoiesis (CH; MIM 162800) and severe congenital neutropenia 1 (SCN1; MIM 202700). Interestingly, the nsSNP I104N, which is known to play a role in causing CH, is predicted here to be located at a protein-protein interface.

## 5.4 Prediction of edgetic effects of SNPs affecting specific pathways

Crosstalk in cellular pathways provides the cell with a robust network of interactions to respond to stimulus. The description of these pathway crosstalk events at molecular level and the mutations that may affect them would open multiple applications in biomedicine, from understanding the homeostatic response of a given drug in a particular population to discovering new personalized scenarios for drug repurposing (Guney et al. 2016; Jaeger, Duran-Frigola, and Aloy 2015). The structural characterization of missense mutations by combining complex structures and docking predictions, as shown in this work, can be essential to achieve this understanding at interactomic level. As an example, structural analysis of the TNNC1 interaction network in MHC phenotype (Figure 6) shows that different nsSNPs could affect the interaction with different proteins. Indeed, mutations affecting TNNT1 binding are in different region than those affecting TNNI1 and TNNI2. Docking-based predictions can help to understand the structural role of additional nsSNPs that are involved in interactions for which there is no available structural data. For instance, based on the docking models, CDK1 binding has been found to be affected by TNNC1 nsSNPs D145E, G159R and E134D; UBE2C binding is found to be affected by E134D; and RBM15B binding is found to be affected by G159R and E134D (**Figure 30**).

In the case of RASopathies, where several of the network nodes are important interaction hubs, a given disease-associated nsSNP at the interface region might have an edgetic effect by affecting certain specific pathways but not others. Therefore, we examined all the pathways that are probably affected by disease-related nsSNPs using NIP extended, and found around 50 affected interactions with proteins that were involved in 38 different pathways. The network topology

provides hints for the role of these nsSNPs. **Figure 19** shows the interactions predicted to be affected by pathological mutations not previously characterized due to the lack of structural data. Interestingly, we identified 26 different pathways involving proteins whose interaction was affected by pathological mutations predicted to be located at hot-spot residues (**Figure 20**).

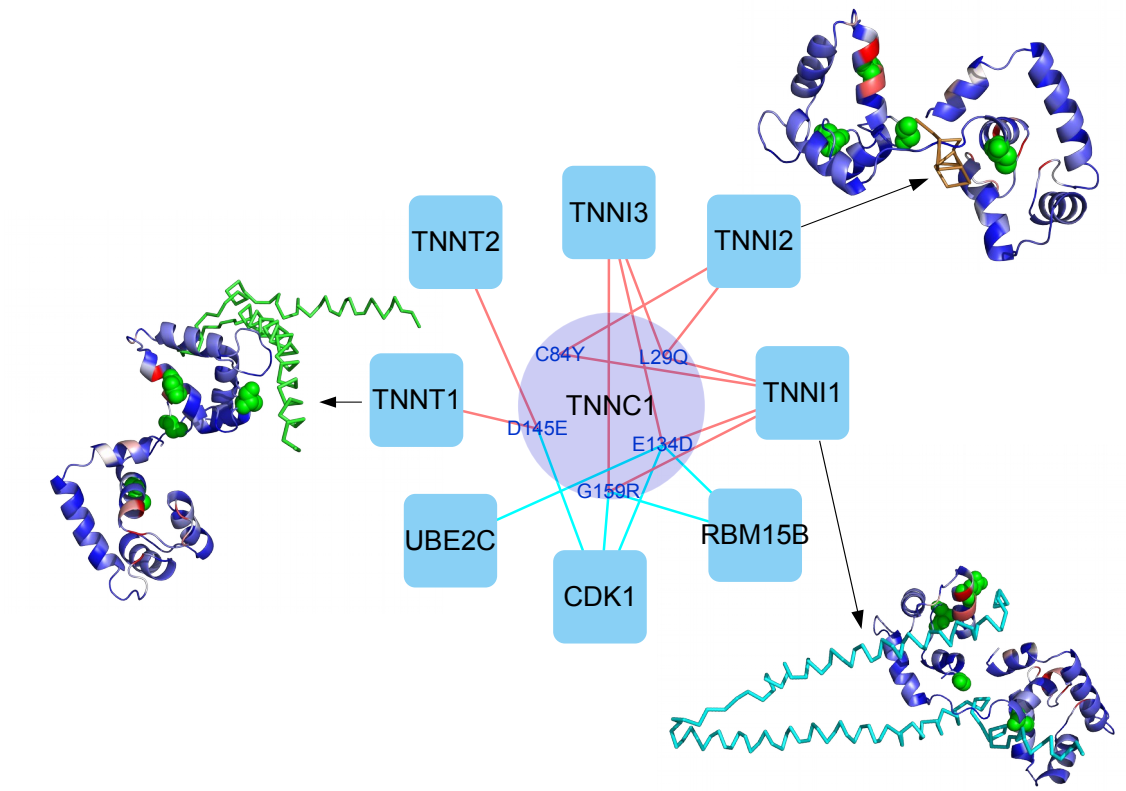



*Figure 30:* **Pathways affected by pathological mutations in RAS/MAPK proteins predicted to be at binding hot-spots**

Proteins of the RAS/MAPK pathway are shown as colored circles, showing pathological mutations that were not previously characterized due to the lack of structural data, but that have been predicted here to be binding hot-spots for docking partner proteins involved in other pathways (linked to the corresponding mutation).

As much as 25 of these pathways are mediated by interaction partners of BRAF and HRAS. The remaining one, the nicotinic acetylcholine receptor signaling pathway, was affected by a pathological mutation in CBL. In total, there are 8 pathways affected by the mutations at the predicted hot-spots that would have not been identified based only on the available structural data **(Figure 20)**. According to our hotspot prediction, the pathways that are involving a larger number of proteins whose interaction was predicted to be affected by pathological mutations are the RAS pathway, VEGF signaling pathway, T cell activation and angiogenesis. All of these pathways

involve interaction partners of both BRAF and HRAS proteins.

## 5.5 Identification and analysis of the protein-protein interactions affected by disease nsSNPs

We previously designed a workflow that allows analyzing the role of nsSNPs on the interaction networks for specific diseases. Here, we aimed to extend our methodology to larger interaction sets, considering the current amount of sequencing projects that are analyzing hundreds to thousands of proteins, in which thousand to millions of variants are identified. As previously described, FFT-based docking programs are computationally cheap and fast, which makes it possible to perform high-throughput docking over the entire high-confidence human interactome. By validating our method with ZDOCK, we had two widely-used docking programs to obtain the normalized interface propensities (NIP) values that are used to identify interface and hot-spot residues. In this way, we had three different but complementary sources of NIP values: the standard pyDock approach (FTDock docking poses scored by pyDock), the ZDOCK docking poses with default scoring, and the ZDOCK docking poses rescored by pyDock. Our method was tested in all the interactions that have a 3D structure available, where we found that the method could place the nsSNPs at the interface regions with high sensitivity and reasonable precision. Applying this methodology to the entire human interactome, we found more than 1,200 interaction affected by a disease-associated nsSNP using the combined information from the available structures and the docking models. These interactions could be potentially new pharmacological targets in a wide variety of diseases. Using extensive interaction networks, we can search for those proteins that create or receive a major burden on the cell with their altered interactions. Thus, we analyzed the interaction network of proteins with the different cellular pathways. With this analysis of the cellular pathways affected by the interactions, we gained complementary knowledge of the proteins that can drive the cell to a disease state. Both ways to analyze the interaction networks produced complementary results. On the one side, in the analysis of direct interactions, we observed the burden over some particular proteins with no nsSNPs at the interface. From the point of view of the pathway analysis, we observed important proteins that are in a crossroad of different pathways. Connecting pathways in this way provide hints of the probable meaningful interactions for the pathway crosstalk. For instance, the CCRK pathway has been reported to possibly play a role in several digestive disorders (Tripathi et al. 2015). Overall, this shows that large-scale computational docking-based calculations could complement current efforts to characterize disease-associated

genetic variants that alter protein sequence (missense mutations), by providing a structural and energetics analysis for many of them. The information generated can be integrated into different mutation pathogenicity predictors or could be used to guide virtual screening experiments to identify small-molecule ligands capable of modulating specific interactions. In both cases, these analyzes aim to have future impact on personalized medicine, helping to improve diagnosis from genetic information, and to develop new therapeutic approaches targeted to individual patients with specific variants.

## 5.6 Future directions

To further improve the predictive capabilities of this approach, we integrate ab initio docking and template-based modeling. It has been recently proposed that currently available structural data on protein-protein complexes is sufficient to provide templates for all protein-protein interfaces, providing that the interacting proteins have structure or a good model (Kundrotas et al. 2012). The problem is that for remote homologous, the existing templates (if any) cannot be directly used to model a protein-protein complex structure with reliability (Negroni, Mosca, and Aloy 2014). Integration of this data with ab initio docking could help to model a larger number of complexes and thus broaden the study of edgetic alteration of an entire disease interaction network.

On the other side, in this thesis we were able to locate disease nsSNPs at protein-protein interfaces for many cases for which there were no structural data available on the interaction. These newly characterized nsSNPs that cause a disease due to their direct involvement in the interface would be interesting targets for identifying small-molecule compounds, for the modulation of signaling cascades or drug development. We have a distinct advantage over other methods that need the 3D structure of the complex to find hot-spots (Oliva and Fernandez-Fuentes 2015), or where templates are necessary to model the protein complex structure (Tuncbag et al. 2011). The major advantages of our method are that it is not computationally expensive, it can identify hot spots and interfaces fairly accurately without prior knowledge of the 3D structure of the complex, and it is an excellent complement for the existing experimental data (it can be optimally combined with SAXS data, mutational experiments, etc.). In addition, including the network analysis facilitates the identification of neglected proteins that participate in the development of disease phenotypes, and opens the possibility of high-throughput docking experiments to characterize specific disease interactomes.

“If you thought that science was certain - well,

that is just an error on your part.”

- Richard Feynman

# Chapter 6 Conclusions

1.- From a systematic analysis of the performance of 73 known functions for the scoring of rigid-body docking poses generated with different docking methods on a standard protein-protein docking benchmark, we found that some of the scoring functions have much better predictive rates than the original functions used in each method

2.- A few scoring functions were sufficiently robust to different types of docking methods, which can be of interest when evaluating a heterogeneous pool of docking models generated by a variety of methods. The combination of different scoring functions looks promising to obtain better predictive rates, but this should be carefully done in order to avoid overtraining.

3.- Integrating scoring functions using methods originally developed for information retrieval and electoral voting provides a powerful method for enhancing the atomic modeling of protein complexes in a way that is tailored to the technique used to generate the models. We have implemented this approach in public available rigid-body protein-protein programs pyDock, ZDOCK, SDOCK.

4.- We have presented here a procedure to improve the characterization of genomic variants involved in protein-protein interactions, especially in cases with low or limited structural information on the binding complexes.

5.- This procedure overcomes current structural data limitations and can help to understand the structural and functional role of genomic variants involved in protein-protein interactions, as well as their edgetic effect on specific protein interaction networks within a given disease.

“Heroes and scholars represent the opposite extremes... The scholar struggles for the benefit of all humanity, sometimes to reduce physical effort, sometimes to reduce pain, and sometimes to postpone death, or at least render it more bearable. In contrast, the patriot sacrifices a rather substantial part of humanity for the sake of his own prestige. His statue is always erected on a pedestal of ruins and corpses... In contrast, all humanity crowns a scholar, love forms the pedestal of his statues, and his triumphs defy the desecration of time and the judgment of history.”

— Santiago Ramón y Cajal, Advice for a Young Investigator

# Chapter 7 References

1000 Genomes Project Consortium, Adam Auton, Lisa D. Brooks, Richard M. Durbin, Erik P. Garrison, Hyun Min Kang, Jan O. Korbel, et al. 2015. “A Global Reference for Human Genetic Variation.” *Nature* 526 (7571): 68–74.

1000 Genomes Project Consortium, Gonçalo R. Abecasis, David Altshuler, Adam Auton, Lisa D. Brooks, Richard M. Durbin, Richard A. Gibbs, Matt E. Hurles, and Gil A. McVean. 2010. “A Map of Human Genome Variation from Population-Scale Sequencing.” *Nature* 467 (7319): 1061–73.

Adzhubei, Ivan A., Steffen Schmidt, Leonid Peshkin, Vasily E. Ramensky, Anna Gerasimova, Peer Bork, Alexey S. Kondrashov, and Shamil R. Sunyaev. 2010. “A Method and Server for Predicting Damaging Missense Mutations.” Nature Methods 7 (4): 248–49.

Afsar Minhas, Fayyaz ul Amir, Fayyaz ul Amir, Brian J. Geiss, and Ben-Hur Asa. 2013. “PAIRpred: Partner-Specific Prediction of Interacting Residues from Sequence and Structure.” Proteins: Structure, Function, and Bioinformatics 82 (7): 1142–55.

Albert, R., H. Jeong, and A. L. Barabasi. 2000. “Error and Attack Tolerance of Complex Networks.” *Nature* 406 (6794): 378–82.

Al-Haggar, Mohammad, Agnieszka Madej-Pilarczyk, Lukasz Kozlowski, Janusz M Bujnicki, Sohier Yahia, Dina Abdel-Hadi, Amany Shams, Nermin Ahmad, Sahar Hamed, and Monika Puzianowska-Kuznicka. 2012. “A Novel Homozygous p.Arg527Leu LMNA Mutation in Two Unrelated Egyptian Families Causes Overlapping Mandibuloacral Dysplasia and Progeria Syndrome.” *European Journal of Human Genetics* 20 (11): 1134–40. doi:10.1038/ejhg.2012.77.

Amberger, J., C.A. Bocchini, A.F. Scott, and A. Hamosh. 2009. “McKusick’s Online Mendelian Inheritance in Man (OMIM).” *Nucleic Acids Res.* 37: D793–96. doi:10.1093/nar/gkn665.

Andrusier, Nelly, Ruth Nussinov, and Haim J. Wolfson. 2007. “FireDock: Fast Interaction Refinement in Molecular Docking.” *Proteins* 69 (1): 139–59.

Anishchenko, Ivan, Petras J. Kundrotas, Alexander V. Tuzikov, and Ilya A. Vakser. 2014. “Protein Models: The Grand Challenge of Protein Docking.” *Proteins* 82 (2): 278–87.

Arnau, Vicente, Sergio Mars, and Ignacio Marín. 2005. “Iterative Cluster Analysis of Protein Interaction Data.” Bioinformatics 21 (3): 364–78.

Bahadur, Ranjit Prasad., and Martin Zacharias. 2008. “The Interface of Protein-Protein Complexes: Analysis of Contacts and Prediction of Interactions.” *Cellular and Molecular Life Sciences* 65 (7–

8): 1059–72. doi:10.1007/s00018-007-7451-x.

Bahadur, Ranjit Prasad, Pinak Chakrabarti, Francis Rodier, and Joël Janin. 2004. “A Dissection of Specific and Non-Specific Protein-Protein Interfaces.” *Journal of Molecular Biology* 336 (4): 943–55.

Barabási, Albert-László, and Zoltán N. Oltvai. 2004. “Network Biology: Understanding the Cell’s Functional Organization.” Nature Reviews. Genetics 5 (2): 101–13.

Barratt, Michael J., and Donald E. Frail. 2012. Drug Repositioning: Bringing New Life to Shelved Assets and Existing Drugs. John Wiley & Sons.

Benson, Dennis A., Ilene Karsch-Mizrachi, David J. Lipman, James Ostell, and David L. Wheeler. 2005. “GenBank.” Nucleic Acids Research 33 (Database issue): D34–38.

Bhalla, U. S., and R. Iyengar. 1999. “Emergent Properties of Networks of Biological Signaling Pathways.” *Science* 283 (5400): 381–87.

Bharathy, Narendra, and Taneja Reshma. 2012. “Methylation Muscles into Transcription Factor Silencing.” *Transcription* 3 (5): 215–20.

Bogan, A. A., and K. S. Thorn. 1998. “Anatomy of Hot-Spots in Protein Interfaces.” *J Mol Biol* 280: 1–9.

Boldon, Lauren, Fallon Laliberte, and Li Liu. 2015. “Review of the Fundamental Theories behind Small Angle X-Ray Scattering, Molecular Dynamics Simulations, and Relevant Integrated Application.” Nano Reviews 6 (February): 25661.

Breiman, Leo. 2001. “10.1023/A:1010933404324.” Machine Learning. doi:10.1023/A:1010933404324.

Breitkreutz, B.J. 2008. “The BioGRID Interaction Database: 2008 Update.” *Nucleic Acids Res.* 36: D637–40. doi:10.1093/nar/gkm1001.

Brown, J. B., and Okuno Yasushi. 2012. “Systems Biology and Systems Chemistry: New Directions for Drug Discovery.” Chemistry & Biology 19 (1): 23–28.

Bush, William S., and Jason H. Moore. 2012. “Chapter 11: Genome-Wide Association Studies.” PLoS Computational Biology 8 (12): e1002822.

Calon, Alexandre, Elisa Espinet, Sergio Palomo-Ponce, Daniele V. F. Tauriello, Mar Iglesias, María Virtudes Céspedes, Marta Sevillano, et al. 2012. “Dependency of Colorectal Cancer on a TGF-β-Driven Program in Stromal Cells for Metastasis Initiation.” Cancer Cell 22 (5): 571–84.

Canutescu, Adrian A., Andrew A. Shelenkov, and Roland L. Dunbrack. 2003. “A Graph-Theory Algorithm for Rapid Protein Side-Chain Prediction.” *Protein Science* 12 (9): 2001–14. doi:10.1110/ps.03154503.

Chang, Yoon Soo, Chang-Min Choi, and Jae Cheol Lee. 2016. “Mechanisms of Epidermal Growth Factor Receptor Tyrosine Kinase Inhibitor Resistance and Strategies to Overcome Resistance in Lung Adenocarcinoma.” *Tuberculosis and Respiratory Diseases* 79 (4): 248–56.Chen, Rong, Li Li, and Weng Zhiping. 2003. “ZDOCK: An Initial-Stage Protein-Docking Algorithm.” Proteins: Structure, Function, and Genetics 52 (1): 80–87.

Chatr-Aryamontri, Andrew, Bobby-Joe Breitkreutz, Rose Oughtred, Lorrie Boucher, Sven Heinicke, Daici Chen, Chris Stark, et al. 2015. “The BioGRID Interaction Database: 2015 Update.” Nucleic Acids Research 43 (Database issue): D470–78.

Cheng, T. M., T. L. Blundell, and J. Fernandez-Recio. 2007. “pyDock: Electrostatics and Desolvation for Effective Scoring of Rigid-Body Protein-Protein Docking.” *Proteins* 68 (2): 503–15.

Cho, Young-Rae, Woochang Hwang, Murali Ramanathan, and Aidong Zhang. 2007. “Semantic Integration to Identify Overlapping Functional Modules in Protein Interaction Networks.” BMC Bioinformatics 8 (July): 265.

Chowbina, Sudhir R., Xiaogang Wu, Fan Zhang, Peter M. Li, Ragini Pandey, Harini N. Kasamsetty, and Jake Y. Chen. 2009. “HPD: An Online Integrated Human Pathway Database Enabling Systems Biology Studies.” BMC Bioinformatics 10 Suppl 11 (October): S5.

Chuang, Gwo-Yu, Dima Kozakov, Ryan Brenke, Stephen R. Comeau, and Sandor Vajda. 2008. “DARS (Decoys As the Reference State) Potentials for Protein-Protein Docking.” Biophysical Journal 95 (9): 4217–27.

Clackson, T., and J. Wells. 1995. “A Hot Spot of Binding Energy in a Hormone-Receptor Interface.” *Science* 267 (5196): 383–86. doi:10.1126/science.7529940.

Clerc, Maurice. 2006. Particle Swarm Optimization. Wiley-ISTE.

Cooper, Gregory M., Julie A. Johnson, Taimour Y. Langaee, Hua Feng, Ian B. Stanaway, Ute I. Schwarz, Marylyn D. Ritchie, et al. 2008. “A Genome-Wide Scan for Common Genetic Variants with a Large Influence on Warfarin Maintenance Dose.” *Blood* 112 (4): 1022–27.

Cordovado, S.K., M. Hendrix, C.N. Greene, S. Mochal, M.C. Earley, P.M. Farrell, M. Kharrazi, W.H. Hannon, and P.W. Mueller. 2012. “CFTR Mutation Analysis and Haplotype Associations in CF Patients.” *Molecular Genetics and Metabolism* 105 (2): 249–54. doi:10.1016/j.ymgme.2011.10.013.

Cortes, Corinna, and Vladimir Vapnik. 1995. “Support-Vector Networks.” Machine Learning 20 (3): 273–97.

Creixell, Pau, Erwin M. Schoof, Janine T. Erler, and Rune Linding. 2012. “Navigating Cancer Network Attractors for Tumor-Specific Therapy.” Nature Biotechnology 30 (9): 842–48.

Crespo, Isaac, Abhimanyu Krishna, Antony Le Béchec, and Antonio del Sol. 2013. “Predicting Missing Expression Values in Gene Regulatory Networks Using a Discrete Logic Modeling Optimization Guided by Network Stable States.” Nucleic Acids Research 41 (1): e8.

Crick, Francis, . 1970. “Central Dogma of Molecular Biology.” *Nature* 227 (5258): 561–63.

Croft, David,. 2010. “Reactome: A Database of Biological Pathways.” Nature Precedings. doi:10.1038/npre.2010.5025.1.

Cui, J., P. Li, G. Li, F. Xu, C. Zhao, Y. Li, Z. Yang, et al. 2007. “AtPID: Arabidopsis Thaliana Protein Interactome Database an Integrative Platform for Plant Systems Biology.” Nucleic Acids Research 36 (Database): D999–1008.

Cukuroglu, Engin, Attila Gursoy, Ruth Nussinov, and Ozlem Keskin. 2014. “Non-Redundant Unique Interface Structures as Templates for Modeling Protein Interactions.” PloS One 9 (1): e86738.

Daelemans, Walter, Hoste Véronique, Fien De Meulder, and Naudts Bart. 2003. “Combined Optimization of Feature Selection and Algorithm Parameters in Machine Learning of Language.” In Lecture Notes in Computer Science, 84–95.

David, Alessia, and Michael J. E. Sternberg. 2015. “The Contribution of Missense Mutations in Core and Rim Residues of Protein–Protein Interfaces to Human Disease.” Journal of Molecular Biology 427 (17): 2886–98.

David, Alessia, Rozami Razali, Mark N. Wass, and Michael J. E. Sternberg. 2012. “Protein-Protein Interaction Sites Are Hot Spots for Disease-Associated Nonsynonymous SNPs.” *Human Mutation* 33 (2): 359–63. doi:10.1002/humu.21656.

Degroeve, Sven, Bernard De Baets, Yves Van de Peer, and Pierre Rouzé. 2002. “Feature Subset Selection for Splice Site Prediction.” Bioinformatics 18 Suppl 2: S75–83.

Derkatch, Irina L., and Susan W. Liebman. 2007. “Prion-Prion Interactions.” *Prion* 1 (3): 161–69.

Díaz-Uriarte, Ramón, and Sara Alvarez de Andrés. 2006. “Gene Selection and Classification of Microarray Data Using Random Forest.” BMC Bioinformatics 7 (January): 3.

Ding, Li, Timothy J. Ley, David E. Larson, Christopher A. Miller, Daniel C. Koboldt, John S. Welch, et al. 2012. “Clonal Evolution in Relapsed Acute Myeloid Leukaemia Revealed by Whole-Genome Sequencing.” *Nature* 481 (7382): 506–10.

Dominguez, C., R. Boelens, and A. M. Bonvin. 2003. “HADDOCK: A Protein-Protein Docking

Approach Based on Biochemical or Biophysical Information." *J Am Chem Soc* 125 (7): 1731–37.

Dong, Qiwen, Wang Xiaolong, Lin Lei, and Guan Yi. 2007. "Exploiting Residue-Level and Profile-Level Interface Propensities for Usage in Binding Sites Prediction of Proteins." BMC Bioinformatics 8 (1): 147.

Elmlund, Dominika, and Elmlund Hans. 2015. "Cryogenic Electron Microscopy and Single-Particle Analysis." Annual Review of Biochemistry 84 (1): 499–517.

Enright, A. J., S. Van Dongen, and C. A. Ouzounis. 2002. "An Efficient Algorithm for Large-Scale Detection of Protein Families." Nucleic Acids Research 30 (7): 1575–84.

Esquivel-Rodriguez, Juan, Vianney Filos-Gonzalez, Bin Li, and Daisuke Kihara. 2014. "Pairwise and Multimeric Protein-Protein Docking Using the LZerD Program Suite." Methods in Molecular Biology 1137: 209–34.

Ewing, Rob M., Chu Peter, Elisma Fred, Li Hongyan, Taylor Paul, Climie Shane, Mcbroom-Cerajewski Linda, et al. 2007. "Large-Scale Mapping of Human Protein–protein Interactions by Mass Spectrometry." Molecular Systems Biology 3. doi:10.1038/msb4100134.

Faure, Guilhem, Jessica Andreani, and Raphaël Guerois. 2012. "InterEvol Database: Exploring the Structure and Evolution of Protein Complex Interfaces." Nucleic Acids Research 40 (Database issue): D847–56.

Feng, Y., A. Kloczkowski, and R. L. Jernigan. 2010. "Potentials 'R' Us Web-Server for Protein Energy Estimations with Coarse-Grained Knowledge-Based Potentials." *BMC Bioinformatics* 11: 92.

Fernandez-Recio, J., M. Totrov, and R. Abagyan. 2003. "ICM-DISCO Docking by Global Energy Optimization with Fully Flexible Side-Chains." *Proteins* 52 (1): 113–17.

Fields, Stanley, Fields Stanley, and Song Ok-kyu. 1989. "A Novel Genetic System to Detect Protein-protein Interactions." *Nature* 340 (6230): 245–46.

Freedman, Matthew L., Alvaro N. A. Monteiro, Simon A. Gayther, Gerhard A. Coetzee, Risch Angela, Plass Christoph, Casey Graham, et al. 2011. "Principles for the Post-GWAS Functional Characterization of Cancer Risk Loci." Nature Genetics 43 (6): 513–18.

Gabb, H. A., R. M. Jackson, and M. J. Sternberg. 1997. "Modelling Protein Docking Using Shape Complementarity, Electrostatics and Biochemical Information." *J Mol Biol* 272: 106–20.

Gallego, P., R. O. Sant'anna, S. Ventura, and D. Reverter. 2016. "Structure of Human Transthyretin in Complex with Tolcapone." doi:10.2210/pdb4d7b/pdb.

Gao, Mu, Zhou Hongyi, and Skolnick Jeffrey. 2015. “Insights into Disease-Associated Mutations in the Human Proteome through Protein Structural Analysis.” Structure 23 (7): 1362–69.

Gerlinger, Marco, Andrew J. Rowan, Horswell Stuart, Larkin James, Endesfelder David, Gronroos Eva, et al. 2012. “Intratumor Heterogeneity and Branched Evolution Revealed by Multiregion Sequencing.” The New England Journal of Medicine 366 (10): 883–92.

Gianni, Stefano, Dogan Jakob, and Jemth Per. 2014. “Distinguishing Induced Fit from Conformational Selection.” Biophysical Chemistry 189: 33–39.

Giot, L., J. S. Bader, C. Brouwer, A. Chaudhuri, B. Kuang, Y. Li, Y. L. Hao, et al. 2003. “A Protein Interaction Map of Drosophila Melanogaster.” *Science* 302 (5651): 1727–36.

Glazko, Galina V., and Frank Emmert-Streib. 2009. “Unite and Conquer: Univariate and Multivariate Approaches for Finding Differentially Expressed Gene Sets.” Bioinformatics 25 (18): 2348–54.

Goh, K-I, M. E. Cusick, D. Valle, B. Childs, M. Vidal, and Barabasi A.-L. 2007. “The Human Disease Network.” Proceedings of the National Academy of Sciences 104 (21): 8685–90.

Goh, Kwang-Il, and In-Geol Choi. 2012. “Exploring the Human Diseasome: The Human Disease Network.” Briefings in Functional Genomics 11 (6): 533–42.

Goodacre, Norman, Nathan Edwards, Mark Danielsen, Peter Uetz, and Cathy Wu. 2016. “Predicting nsSNPs the Disrupt Protein-Protein Interactions Using Docking.” IEEE/ACM Transactions on Computational Biology and Bioinformatics / IEEE, ACM, January. doi:10.1109/TCBB.2016.2520931.

Grosdidier, Solène, and Juan Fernández-Recio. 2008. “Identification of Hot-Spot Residues in Protein-Protein Interactions by Computational Docking.” *BMC Bioinformatics* 9 (1): 447. doi:10.1186/1471-2105-9-447.

Guruharsha, K. G., Jean-François Rual, Bo Zhai, Julian Mintseris, Pujita Vaidya, Namita Vaidya, Chapman Beekman, et al. 2011. “A Protein Complex Network of Drosophila Melanogaster.” *Cell* 147 (3): 690–703.

Halperin, Inbal, Halperin Inbal, Ma Buyong, Wolfson Haim, and Nussinov Ruth. 2002. “Principles of Docking: An Overview of Search Algorithms and a Guide to Scoring Functions.” Proteins: Structure, Function, and Genetics 47 (4): 409–43.

Hamp, Tobias, and Rost Burkhard. 2012. “Alternative Protein-Protein Interfaces Are Frequent Exceptions.” PLoS Computational Biology 8 (8): e1002623.

Hartuv, Erez, and Shamir Ron. 2000. “A Clustering Algorithm Based on Graph Connectivity.”

Information Processing Letters 76 (4-6): 175–81.

Hartwell, L. H., J. J. Hopfield, S. Leibler, and A. W. Murray. 1999. “From Molecular to Modular Cell Biology.” *Nature* 402 (6761 Suppl): C47–52.

Higgins, Desmond G., and Paul M. Sharp. 1988. “CLUSTAL: A Package for Performing Multiple Sequence Alignment on a Microcomputer.” *Gene* 73 (1): 237–44.

Hou, Yong, Song Luting, Zhu Ping, Zhang Bo, Tao Ye, Xu Xun, et al. 2012. “Single-Cell Exome Sequencing and Monoclonal Evolution of a JAK2-Negative Myeloproliferative Neoplasm.” *Cell* 148 (5): 873–85.

Hu, Z., B. Ma, H. Wolfson, and R. Nussinov. 2000. “Conservation of Polar Residues as Hot-Spots at Protein Interfaces.” *Proteins* 39: 331–42.

Huang, Hui, Xiaogang Wu, Madhankumar Sonachalam, Sammed N. Mandape, Ragini Pandey, Karl F. MacDorman, Ping Wan, and Jake Y. Chen. 2012. “PAGED: A Pathway and Gene-Set Enrichment Database to Enable Molecular Phenotype Discoveries.” BMC Bioinformatics 13 Suppl 15 (September): S2.

Huang, Sheng-You. 2014. “Search Strategies and Evaluation in Protein–protein Docking: Principles, Advances and Challenges.” Drug Discovery Today 19 (8): 1081–96.

Hughey, Jacob J., Trevor Hastie, and Atul J. Butte. 2016. “ZeitZeiger: Supervised Learning for High-Dimensional Data from an Oscillatory System.” Nucleic Acids Research, January. doi:10.1093/nar/gkw030.

Hwang, H., T. Vreven, J. Janin, and Z. Weng. 2010. “Protein-Protein Docking Benchmark Version 4.0.” *Proteins* 78 (15): 3111–14.

Hwang, Howook, Thom Vreven, and Zhiping Weng. 2014. “Binding Interface Prediction by Combining Protein-Protein Docking Results.” *Proteins* 82 (1): 57–66.

Ideker, T., V. Thorsson, J. A. Ranish, R. Christmas, J. Buhler, J. K. Eng, R. Bumgarner, D. R. Goodlett, R. Aebersold, and L. Hood. 2001. “Integrated Genomic and Proteomic Analyses of a Systematically Perturbed Metabolic Network.” *Science* 292 (5518): 929–34.

International HapMap 3 Consortium, David M. Altshuler, Richard A. Gibbs, Leena Peltonen, David M. Altshuler, Richard A. Gibbs, Leena Peltonen, et al. 2010. “Integrating Common and Rare Genetic Variation in Diverse Human Populations.” *Nature* 467 (7311): 52–58.

International HapMap Consortium. 2005. “A Haplotype Map of the Human Genome.” *Nature* 437 (7063): 1299–1320.

International Human Genome Sequencing Consortium. 2004. “Finishing the Euchromatic Sequence

of the Human Genome.” *Nature* 431 (7011): 931–45.

Ito, T., K. Tashiro, S. Muta, R. Ozawa, T. Chiba, M. Nishizawa, K. Yamamoto, S. Kuhara, and Y. Sakaki. 2000. “Toward a Protein-Protein Interaction Map of the Budding Yeast: A Comprehensive System to Examine Two-Hybrid Interactions in All Possible Combinations between the Yeast Proteins.” Proceedings of the National Academy of Sciences 97 (3): 1143–47.

Jaeger, Samira, and Aloy Patrick. 2012. “From Protein Interaction Networks to Novel Therapeutic Strategies.” IUBMB Life 64 (6): 529–37.

Jaeger, Samira, Miquel Duran-Frigola, and Patrick Aloy. 2015. “Drug Sensitivity in Cancer Cell Lines Is Not Tissue-Specific.” *Molecular Cancer* 14 (1): 40. doi:10.1186/s12943-015-0312-6.

Janes, Kevin A., John G. Albeck, Suzanne Gaudet, Peter K. Sorger, Douglas A. Lauffenburger, and Michael B. Yaffe. 2005. “A Systems Model of Signaling Identifies a Molecular Basis Set for Cytokine-Induced Apoptosis.” *Science* 310 (5754): 1646–53.

Janin, Joël, Kim Henrick, John Moult, Lynn Ten Eyck, Michael J. E. Sternberg, Sandor Vajda, Ilya Vakser, Shoshana J. Wodak, and Critical Assessment of PRedicted Interactions. 2003. “CAPRI: A Critical Assessment of PRedicted Interactions.” *Proteins* 52 (1): 2–9.

Janjić, Vuk, and Pržulj Nataša. 2012. “The Core Diseasome.” Molecular bioSystems 8 (10): 2614.

Jiménez-García, Brian, Carles Pons, Dmitri I. Svergun, Pau Bernadó, and Juan Fernández-Recio. 2015. “pyDockSAXS: Protein-Protein Complex Structure by SAXS and Computational Docking.” Nucleic Acids Research 43 (W1): W356–61.

Joachims, Thorsten, Finley Thomas, and Chun-Nam John Yu. 2009. “Cutting-Plane Training of Structural SVMs.” Machine Learning 77 (1): 27–59.

Joachims, Thorsten. 2005. “A Support Vector Method for Multivariate Performance Measures.” In *Proceedings of the 22nd International Conference on Machine Learning*, 377–84. ACM.

Joachims, Thorsten. 2006. “Training Linear SVMs in Linear Time.” In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 217–26. ACM.

Kandpal, Raj, Saviola Beatrice, and Felton Jeffrey. 2009. “The Era of ’Omics Unlimited.” *BioTechniques* 46 (5): 351–55.

Kastritis, P. L., I. H. Moal, H. Hwang, Z. Weng, P. A. Bates, A. M. Bonvin, and J. Janin. 2011. “A Structure-Based Benchmark for Protein-Protein Binding Affinity.” *Protein Sci* 20 (3): 482–91.

Kastritis, Panagiotis, Iain H. Moal, Hwang Howook, Weng Zhiping, Paul A. Bates, Alexandre M Bonvin, and Janin Joël. 2011. “A Structure-Based Benchmark for Protein-Protein Binding

Affinity.” Protein Science: A Publication of the Protein Society 20 (3): 482–91.

Katchalski-Katzir, E., I. Shariv, M. Eisenstein, A. A. Friesem, C. Aflalo, and I. A. Vakser. 1992. “Molecular Surface Recognition: Determination of Geometric Fit between Proteins and Their Ligands by Correlation Techniques.” Proceedings of the National Academy of Sciences of the United States of America 89 (6): 2195–99.

Katsonis, Panagiotis, and Olivier Lichtarge. 2014. “A Formal Perturbation Equation between Genotype and Phenotype Determines the Evolutionary Action of Protein-Coding Variations on Fitness.” Genome Research 24 (12): 2050–58.

Keshava Prasad, T.S. 2009. “Human Protein Reference Database-2009 Update.” *Nucleic Acids Res.* 37: D767–72. doi:10.1093/nar/gkn892.

Keskin, Ozlem, and Ruth Nussinov. 2007. “Similar Binding Sites and Different Partners: Implications to Shared Proteins in Cellular Pathways.” *Structure* 15 (3): 341–54. doi:10.1016/j.str.2007.01.007.

Khatri, Purvesh, Marina Sirota, and Atul J. Butte. 2012. “Ten Years of Pathway Analysis: Current Approaches and Outstanding Challenges.” PLoS Computational Biology 8 (2): e1002375.

Kiel, C., and L. Serrano. 2014. “Structure-Energy-Based Predictions and Network Modelling of RASopathy and Cancer Missense Mutations.” *Molecular Systems Biology* 10 (5): 727–727. doi:10.1002/msb.20145092.

Kiel, Christina, and Luis Serrano. 2014. 2014. “Structure-Energy-Based Predictions and Network Modelling of RASopathy and Cancer Missense Mutations.” *Molecular Systems Biology* 10 (5): 727–727. doi:10.1002/msb.20145092.

Kiel, Christina, Pedro Beltrao, and Luis Serrano. 2008. “Analyzing Protein Interaction Networks Using Structural Information.” Annual Review of Biochemistry 77: 415–41.

King, A. D., N. Przulj, and I. Jurisica. 2004. “Protein Complex Prediction via Cost-Based Clustering.” Bioinformatics 20 (17): 3013–20.

Kitano, Hiroaki. 2004. “Biological Robustness.” *Nature Reviews Genetics* 5 (11): 826–37. doi:10.1038/nrg1471.

Kitano, Hiroaki. 2004. “Opinion: Cancer as a Robust System: Implications for Anticancer Therapy.” Nature Reviews. Cancer 4 (3): 227–35.

Koepfli, Klaus-Peter, Koepfli Klaus-Peter, Paten Benedict, Stephen J. O’Brien, and The Genome 10k Community. 2015. “The Genome 10K Project: A Way Forward.” Annual Review of Animal Biosciences 3 (1): 57–111.

Kozakov, Dima, Brenke Ryan, Stephen R. Comeau, and Vajda Sandor. 2006. “PIPER: An FFT-Based Protein Docking Program with Pairwise Potentials.” Proteins: Structure, Function, and Bioinformatics 65 (2): 392–406.

Kryshtafovych, Andriy, Fidelis Krzysztof, and Moult John. 2013. “CASP10 Results Compared to Those of Previous CASP Experiments.” Proteins: Structure, Function, and Bioinformatics 82: 164–74.

Kundrotas, P. J., Z. Zhu, J. Janin, and I. A. Vakser. 2012. “Templates Are Available to Model Nearly All Complexes of Structurally Characterized Proteins.” *Proceedings of the National Academy of Sciences* 109 (24): 9438–41. doi:10.1073/pnas.1200678109.

Kuntz, I. D., J. M. Blaney, S. J. Oatley, R. Langridge, and T. E. Ferrin. 1982. “A Geometric Approach to Macromolecule-Ligand Interactions.” Journal of Molecular Biology 161 (2): 269–88.

Kuser, Paula, Cupri Fabio, Bleicher Lucas, and Polikarpov Igor. 2008. “Crystal Structure of Yeast Hexokinase PI in Complex with Glucose: A Classical ‘induced Fit’ Example Revised.” Proteins: Structure, Function, and Bioinformatics 72 (2): 731–40.

Lander, Eric S. 2011. “Initial Impact of the Sequencing of the Human Genome.” *Nature* 470 (7333): 187–97.

Lemmon, Mark A., and Joseph Schlessinger. 2010. “Cell Signaling by Receptor Tyrosine Kinases.” *Cell* 141 (7): 1117–34.

Lensink, Marc F., and Shoshana J. Wodak. 2014. “Score_set: A CAPRI Benchmark for Scoring Protein Complexes: A Benchmark for Scoring Protein Complexes.” *Proteins: Structure, Function, and Bioinformatics* 82 (11): 3163–69. doi:10.1002/prot.24678.

Lensink, Marc F., Raúl Méndez, and Shoshana J. Wodak. 2007. “Docking and Scoring Protein Complexes: CAPRI 3rd Edition.” *Proteins: Structure, Function, and Bioinformatics* 69 (4): 704–18. doi:10.1002/prot.21804.

Levy, Emmanuel D. 2010. “A Simple Definition of Structural Regions in Proteins and Its Use in Analyzing Interface Evolution.” Journal of Molecular Biology 403 (4): 660–70.

Levy, Emmanuel D., Jose B. Pereira-Leal, Cyrus Chothia, and Sarah A. Teichmann. 2006. “3D Complex: A Structural Classification of Protein Complexes.” PLoS Computational Biology 2 (11): e155.

Li, Chen Rong, and Weng Zhiping. 2003. “RDOCK: Refinement of Rigid-Body Protein Docking Predictions.” Proteins: Structure, Function, and Genetics 53 (3): 693–707.

Li, Guipeng, Ming Li, Yiwei Zhang, Dong Wang, Rong Li, Roger Guimerà, Juntao Tony Gao, and

Michael Q. Zhang. 2014. “ModuleRole: A Tool for Modulization, Role Determination and Visualization in Protein-Protein Interaction Networks.” PloS One 9 (5): e94608.

Li, Siming, Christopher M. Armstrong, Nicolas Bertin, Hui Ge, Stuart Milstein, Mike Boxem, Pierre-Olivier Vidalain, et al. 2004. “A Map of the Interactome Network of the Metazoan C. Elegans.” *Science* 303 (5657): 540–43.

Lilley, David M. J., and Eckstein Fritz. n.d. “Chapter 1. Ribozymes and RNA Catalysis: Introduction and Primer.” In Ribozymes and RNA Catalysis, 1–10.

Liu, S., and I. A. Vakser. 2011. “DECK: Distance and Environment-Dependent, Coarse-Grained, Knowledge-Based Potentials for Protein-Protein Docking.” *BMC Bioinformatics* 12: 280.

Liu, Shiyong, Ying Gao, and Ilya A. Vakser. 2008. “DOCKGROUND Protein-Protein Docking Decoy Set.” Bioinformatics 24 (22): 2634–35.

Lo, Yu-Shu, Chen Yung-Chiang, and Yang Jinn-Moon. 2010. “3D-Interologs: An Evolution Database of Physical Protein- Protein Interactions across Multiple Genomes.” BMC Genomics 11 (Suppl 3): S7.

Lord, P. W., R. D. Stevens, A. Brass, and C. A. Goble. 2003. “Investigating Semantic Similarity Measures across the Gene Ontology: The Relationship between Sequence and Annotation.” Bioinformatics 19 (10): 1275–83.

Luciani, Davide, and Gianfranco Bazzoni. 2012. “From Networks of Protein Interactions to Networks of Functional Dependencies.” BMC Systems Biology 6 (May): 44.

Ma’ayan, Avi. 2009. “Insights into the Organization of Biochemical Regulatory Networks Using Graph Theory Analyses.” The Journal of Biological Chemistry 284 (9): 5451–55.

MacBeath, Gavin, 2002. “Protein Microarrays and Proteomics.” Nature Genetics 32 (Supp): 526–32.

Maciag, Karolina, Steven J. Altschuler, Michael D. Slack, Nevan J. Krogan, Andrew Emili, Jack F. Greenblatt, Tom Maniatis, and Lani F. Wu. 2006. “Systems-Level Analyses Identify Extensive Coupling among Gene Expression Machines.” Molecular Systems Biology 2 (January): 2006.0003.

Maere, Steven, Karel Heymans, and Martin Kuiper. 2005. “BiNGO: A Cytoscape Plugin to Assess Overrepresentation of Gene Ontology Categories in Biological Networks.” *Bioinformatics* 21 (16): 3448–49.

Marion, Dominique. 2013. “An Introduction to Biological NMR Spectroscopy.” Molecular & Cellular Proteomics: MCP 12 (11): 3006–25.

Martin, Juliette, and Richard Lavery. 2012. “Arbitrary Protein−protein Docking Targets Biologically Relevant Interfaces.” *BMC Biophysics* 5 (1): 7. doi:10.1186/2046-1682-5-7.

Marx, Vivien, . 2013. “Biology: The Big Challenges of Big Data.” *Nature* 498 (7453): 255–60.

Mashiach, Efrat, Ruth Nussinov, and Haim J. Wolfson. 2010. “FiberDock: Flexible Induced-Fit Backbone Refinement in Molecular Docking.” *Proteins* 78 (6): 1503–19.

Matthews, L. R. 2001. “Identification of Potential Interaction Networks Using Sequence-Based Searches for Conserved Protein-Protein Interactions or ‘Interologs.’” Genome Research 11 (12): 2120–26.

Mering, C. v.. 2003. “STRING: A Database of Predicted Functional Associations between Proteins.” Nucleic Acids Research 31 (1): 258–61.

Mi, H. 2004. “The PANTHER Database of Protein Families, Subfamilies, Functions and Pathways.” *Nucleic Acids Research* 33 (Database issue): D284–88. doi:10.1093/nar/gki078.

Moal, I. H., and Juan Fernández-Recio. 2012. “SKEMPI: A Structural Kinetic and Energetic Database of Mutant Protein Interactions and Its Use in Empirical Models.” Bioinformatics 28 (20): 2600–2607.

Moal, I. H., and P. A. Bates. 2010. “SwarmDock and the Use of Normal Modes in Protein-Protein Docking.” *Int J Mol Sci* 11 (10): 3623–48.

Moal, I. H., Brian Jiménez-García, and Juan Fernández-Recio. 2015. “CCharPPI Web Server: Computational Characterization of Protein-Protein Interactions from Structure.” *Bioinformatics (Oxford, England)* 31 (1): 123–25. doi:10.1093/bioinformatics/btu594.

Moal, I. H., Mieczyslaw Torchala, P. A. Bates, and Juan Fernandez-Recio. 2013. “The Scoring of Poses in Protein-Protein Docking: Current Capabilities and Future Directions.” *BMC Bioinformatics* 14 (1): 286.

Moorthy, Kohbalan, and Mohd Saberi Mohamad. 2011. “Random Forest for Gene Selection and Microarray Data Classification.” *Bioinformation* 7 (3): 142–46.

Morrison, Kim L., and Gregory A. Weiss. 2001. “Combinatorial Alanine-Scanning.” Current Opinion in Chemical Biology 5 (3): 302–7.

Mosca, Roberto, Arnaud Céol, Amelie Stein, Roger Olivella, and Patrick Aloy. 2014. “3did: A Catalog of Domain-Based Interactions of Known Three-Dimensional Structure.” Nucleic Acids Research 42 (Database issue): D374–79.

Mosca, Roberto, Arnaud Céol, and Patrick Aloy. 2013. “Interactome3D: Adding Structural Details to Protein Networks.” *Nature Methods* 10 (1): 47–53. doi:10.1038/nmeth.2289.

Mosca, Roberto, Carles Pons, Juan Fernández-Recio, and Patrick Aloy. 2009. “Pushing Structural Information into the Yeast Interactome by High-Throughput Protein Docking Experiments.” *PLoS Computational Biology* 5 (8): e1000490

Mosca, Roberto, Jofre Tenorio-Laranga, Roger Olivella, Victor Alcalde, Arnaud Céol, Montserrat Soler-López, and Patrick Aloy. 2015. “dSysMap: Exploring the Edgetic Role of Disease Mutations.” *Nature Methods* 12 (3): 167–68. doi:10.1038/nmeth.3289.

Müller, Arne, Robert M. MacCallum, and Michael J. E. Sternberg. 2002. “Structural Characterization of the Human Proteome.” Genome Research 12 (11): 1625–41.

Myint, Kyaw-Zeyar, Wang Lirong, Tong Qin, and Xie Xiang-Qun. 2012. “Molecular Fingerprint-Based Artificial Neural Networks QSAR for Ligand Biological Activity Predictions.” Molecular Pharmaceutics 9 (10): 2912–23.

Negroni, Jacopo, Roberto Mosca, and Patrick Aloy. 2014. “Assessing the Applicability of Template-Based Protein Docking in the Twilight Zone.” *Structure* 22 (9): 1356–62. doi:10.1016/j.str.2014.07.009.

Nishimura, Darryl, and Nishimura Darryl. 2001. “BioCarta.” Biotech Software & Internet Report 2 (3): 117–20.

Ofran, Y., and B. Rost. 2003. “Analysing Six Types of Protein-Protein Interfaces.” *J Mol Biol* 325 (2): 377–87.

Ogata H, Goto S, Sato K, Fujibuchi W, Bono H, Kanehisa M. KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res. 1999;27: 29–34. doi:10.1093/nar/27.1.29

Oliva, B., and N. Fernandez-Fuentes. 2015. “Knowledge-Based Modeling of Peptides at Protein Interfaces: PiPreD.” *Bioinformatics* 31 (9): 1405–10. doi:10.1093/bioinformatics/btu838.

Park, B., and M. Levitt. 1996. “Energy Functions That Discriminate X-Ray and near Native Folds from Well-Constructed Decoys.” *J Mol Biol* 258 (2): 367–92.

Pierce, B. G. and Zhiping Weng. 2007. “ZRANK: Reranking Protein Docking Predictions with an Optimized Energy Function.” Proteins: Structure, Function, and Bioinformatics 67 (4): 1078–86.

Pierce, B. G., Yuichiro Hourai, and Zhiping Weng. 2011. “Accelerating Protein Docking in ZDOCK Using an Advanced 3D Convolution Library.” Edited by Ozlem Keskin. *PLoS ONE* 6 (9): e24657. doi:10.1371/journal.pone.0024657.

Pizzuti, Clara, and Simona E. Rombo. 2014. “Algorithms and Tools for Protein-Protein Interaction Networks Clustering, with a Special Focus on Population-Based Stochastic Methods.” Bioinformatics 30 (10): 1343–52.

Pokarowski, P., A. Kloczkowski, R. L. Jernigan, N. S. Kothari, M. Pokarowska, and A. Kolinski. 2005. “Inferring Ideal Amino Acid Interaction Forms from Statistical Protein Contact Potentials.” *Proteins* 59: 49–57.

Pons, Carles, D. Talavera, X. de la Cruz, M. Orozco, and J. Fernandez-Recio. 2011. “Scoring by Intermolecular Pairwise Propensities of Exposed Residues (SIPPER): A New Efficient Potential for Protein-Protein Docking.” *J Chem Inf Model* 51 (2): 370–77.

Pons, Carles, Solène Grosdidier, Albert Solernou, Laura Pérez-Cano, and Juan Fernández-Recio. 2010. “Present and Future Challenges and Limitations in Protein-Protein Docking.” *Proteins: Structure, Function, and Bioinformatics* 78 (1): 95–108. doi:10.1002/prot.22564.

Project, Encode, Darryl L. Leja, and Ewan Birney. 2012. The ENCODE Project Encyclopedia of DNA Elements.

Puig, O., F. Caspary, G. Rigaut, B. Rutz, E. Bouveret, E. Bragado-Nilsson, M. Wilm, and B. Séraphin. 2001. “The Tandem Affinity Purification (TAP) Method: A General Procedure of Protein Complex Purification.” Methods 24 (3): 218–29.

Qiu, Yu-Qing. 2013. “KEGG Pathway Database.” In Encyclopedia of Systems Biology, 1068–69.

Redington, Patrick K. 1992. “MOLFIT: A Computer Program for Molecular Superposition.” Computers & Chemistry 16 (3): 217–22.

Ritchie, Marylyn D., Joshua C. Denny, Dana C. Crawford, Andrea H. Ramirez, Justin B. Weiner, Jill M. Pulley, Melissa A. Basford, et al. 2010. “Robust Replication of Genotype-Phenotype Associations across Multiple Diseases in an Electronic Medical Record.” American Journal of Human Genetics 86 (4): 560–72.

Rolland, Thomas, Murat Taşan, Benoit Charloteaux, Samuel J. Pevzner, Quan Zhong, Nidhi Sahni, Song Yi, et al. 2014. “A Proteome-Scale Map of the Human Interactome Network.” *Cell* 159 (5): 1212–26.

Rual, Jean-François, Kavitha Venkatesan, Tong Hao, Tomoko Hirozane-Kishikawa, Amélie Dricot, Ning Li, Gabriel F. Berriz, et al. 2005. “Towards a Proteome-Scale Map of the Human Protein-Protein Interaction Network.” *Nature* 437 (7062): 1173–78.

Russell, Rupert J., Lesley F. Haire, David J. Stevens, Patrick J. Collins, Yi Pu Lin, G. Michael Blackburn, Alan J. Hay, Steven J. Gamblin, and John J. Skehel. 2006. “The Structure of H5N1 Avian Influenza Neuraminidase Suggests New Opportunities for Drug Design.” *Nature* 443 (7107): 45–49.

Schiffer, M., C. Ainsworth, Z. B. Xu, W. Carperos, K. Olsen, A. Solomon, F. J. Stevens, and C. H.

Chang. 1989. “Structure of a Second Crystal Form of Bence-Jones Protein Loc: Strikingly Different Domain Associations in Two Crystal Forms of a Single Protein.” *Biochemistry* 28 (9): 4066–72.

Schneidman-Duhovny, D., Y. Inbar, R. Nussinov, and H. J. Wolfson. 2005. “PatchDock and SymmDock: Servers for Rigid and Symmetric Docking.” Nucleic Acids Research 33 (Web Server): W363–67.

Schulze, Markus. 2011. “A New Monotonic, Clone-Independent, Reversal Symmetric, and Condorcet-Consistent Single-Winner Election Method.” *Social Choice and Welfare* 36 (2): 267–303. doi:10.1007/s00355-010-0475-4.

Scott, Alan F., Amberger Joanna, Brylawski Brandon, and Victor A. McKusick. n.d. “OMIM: Online Mendelian Inheritance in Man.” In Bioinformatics: Databases and Systems, 77–84.

Sevilla, J. L., V. Segura, A. Podhorski, E. Guruceaga, J. M. Mato, L. A. Martinez-Cruz, F. J. Corrales, and A. Rubio. 2005. “Correlation between Gene Expression and GO Semantic Similarity.” IEEE/ACM Transactions on Computational Biology and Bioinformatics / IEEE, ACM 2 (4): 330–38.

Shah, Naishadh, Ayesha Sattar, Michael Benanti, Scott Hollander, and Lanna Cheuck. 2006. “Magnetic Resonance Spectroscopy as an Imaging Tool for Cancer: A Review of the Literature.” The Journal of the American Osteopathic Association 106 (1): 23–27.

Shah, Sohrab P., Andrew Roth, Rodrigo Goya, Arusha Oloumi, Gavin Ha, Yongjun Zhao, Gulisa Turashvili, et al. 2012. “The Clonal and Mutational Evolution Spectrum of Primary Triple-Negative Breast Cancers.” *Nature* 486 (7403): 395–99.

Shoemaker, Benjamin A., Dachuan Zhang, Manoj Tyagi, Ratna R. Thangudu, Jessica H. Fong, Aron Marchler-Bauer, Stephen H. Bryant, Thomas Madej, and Anna R. Panchenko. 2012. “IBIS (Inferred Biomolecular Interaction Server) Reports, Predicts and Integrates Multiple Types of Conserved Interactions for Proteins.” Nucleic Acids Research 40 (Database issue): D834–40.

Sinha, Rohita, Petras J. Kundrotas, and Ilya A. Vakser. 2012. “Protein Docking by the Interface Structure Similarity: How Much Structure Is Needed?” PloS One 7 (2): e31349.

Smyth, M. S. 2000. “X Ray Crystallography.” Molecular Pathology: MP 53 (1): 8–14.

Stelzl, Ulrich, Uwe Worm, Maciej Lalowski, Christian Haenig, Felix H. Brembeck, Heike Goehler, Martin Stroedicke, et al. 2005. “A Human Protein-Protein Interaction Network: A Resource for Annotating the Proteome.” *Cell* 122 (6): 957–68.

Stingele, Silvia, Gabriele Stoehr, Karolina Peplowska, Jürgen Cox, Matthias Mann, and Zuzana

Storchova. 2012. “Global Analysis of Genome, Transcriptome and Proteome Reveals the Response to Aneuploidy in Human Cells.” Molecular Systems Biology 8: 608.

Strogatz, S. H. 2001. “Exploring Complex Networks.” *Nature* 410 (6825): 268–76.

Stumpf, Michael P. H., Thomas Thorne, Eric de Silva, Ronald Stewart, Hyeong Jun An, Michael Lappe, and Carsten Wiuf. 2008. “Estimating the Size of the Human Interactome.” Proceedings of the National Academy of Sciences of the United States of America 105 (19): 6959–64.

Teichmann, S. A. 2002. “Principles of Protein-Protein Interactions.” Bioinformatics 18 (Suppl 2): S249–S249.

Teng, Shaolei, Thomas Madej, Anna Panchenko, and Emil Alexov. 2009. “Modeling Effects of Human Single Nucleotide Polymorphisms on Protein-Protein Interactions.” *Biophysical Journal* 96 (6): 2178–88. doi:10.1016/j.bpj.2008.12.3904.

The UniProt Consortium. 2007. “The Universal Protein Resource (UniProt).” Nucleic Acids Research 36 (Database): D190–95.

Tobi, Dror, 2010. “Designing Coarse Grained-and Atom Based-Potentials for Protein-Protein Docking.” BMC Structural Biology 10 (1): 40.

Tobi, Dror, and Ivet Bahar. 2006. “Optimal Design of Protein Docking Potentials: Efficiency and Limitations.” *Proteins* 62 (4): 970–81.

Tripathi, Sushil, Åsmund Flobak, Konika Chawla, Anaïs Baudot, Torunn Bruland, Liv Thommesen, Martin Kuiper, and Astrid Lægreid. 2015. “The Gastrin and Cholecystokinin Receptors Mediated Signaling Network: A Scaffold for Data Analysis and New Hypotheses on Regulatory Mechanisms.” *BMC Systems Biology* 9 (July): 40.

Tryka, Kimberly A., Luning Hao, Anne Sturcke, Yumi Jin, Zhen Y. Wang, Lora Ziyabari, Moira Lee, et al. 2014. “NCBI’s Database of Genotypes and Phenotypes: dbGaP.” Nucleic Acids Research 42 (Database issue): D975–79.

Tuncbag, Nurcan, Attila Gursoy, Ruth Nussinov, and Ozlem Keskin. 2011. “Predicting Protein-Protein Interactions on a Proteome Scale by Matching Evolutionary and Structural Similarities at Interfaces Using PRISM.” *Nature Protocols* 6 (9): 1341–54. doi:10.1038/nprot.2011.367.

Uetz, P., and R. E. Hughes. 2000. “Systematic and Large-Scale Two-Hybrid Screens.” Current Opinion in Microbiology 3 (3): 303–8.

Vajda, Sandor. 2005. “Classification of Protein Complexes Based on Docking Difficulty.” *Proteins: Structure, Function, and Bioinformatics* 60 (2): 176–80. doi:10.1002/prot.20554.

Vajda, Sandor, and Dima Kozakov. 2009. “Convergence and Combination of Methods in Protein-

Protein Docking.” Current Opinion in Structural Biology 19 (2): 164–70.

Valencia, Alfonso, and Pazos Florencio. 2005. “Prediction of Protein-Protein Interactions from Evolutionary Information.” In Methods of Biochemical Analysis, 409–26.

Vavricka, Christopher J., Qing Li, Yan Wu, Jianxun Qi, Mingyang Wang, Yue Liu, Feng Gao, et al. 2011. “Structural and Functional Analysis of Laninamivir and Its Octanoate Prodrug Reveals Group Specific Mechanisms for Influenza NA Inhibition.” PLoS Pathogens 7 (10): e1002249.

Vazquez, Alexei, Alessandro Flammini, Amos Maritan, and Alessandro Vespignani. 2003. “Global Protein Function Prediction from Protein-Protein Interaction Networks.” Nature Biotechnology 21 (6): 697–700.

Venter, J. Craig, Karin Remington, John F. Heidelberg, Aaron L. Halpern, Doug Rusch, Jonathan A. Eisen, Dongying Wu, et al. 2004. “Environmental Genome Shotgun Sequencing of the Sargasso Sea.” *Science* 304 (5667): 66–74.

Vidal, Marc, Michael E. Cusick, and Albert-László Barabási. 2011. “Interactome Networks and Human Disease.” *Cell* 144 (6): 986–98.

Viswanath, S., D. V. Ravikant, and R. Elber. 2012. “Improving Ranking of Models for Protein Complexes with Side Chain Modeling and Atomic Potentials.” *Proteins* 81 (4): 592–606.

Vreven, Thom, Iain H. Moal, Anna Vangone, Brian G. Pierce, Panagiotis L. Kastritis, Mieczyslaw Torchala, Raphael Chaleil, et al. 2015. “Updates to the Integrated Protein–Protein Interaction Benchmarks: Docking Benchmark Version 5 and Affinity Benchmark Version 2.” *Journal of Molecular Biology*, July. doi:10.1016/j.jmb.2015.07.016.

Wang, Baoying. 2014. Big Data Analytics in Bioinformatics and Healthcare. IGI Global.

Welter, D., J. MacArthur, J. Morales, T. Burdett, P. Hall, H. Junkins, A. Klemm, et al. 2013. “The NHGRI GWAS Catalog, a Curated Resource of SNP-Trait Associations.” Nucleic Acids Research 42 (D1): D1001–6.

Winter, Christof, Andreas Henschel, Wan Kyu Kim, and Michael Schroeder. 2006. “SCOPPI: A Structural Classification of Protein-Protein Interfaces.” Nucleic Acids Research 34 (Database issue): D310–14.

Wu, Xiaogang, and Jake Y. Chen. 2012. “An Evaluation for Merging Signaling Pathways by Using Protein-Protein Interaction Data.” In Proceedings 2012 IEEE International Workshop on Genomic Signal Processing and Statistics (GENSIPS). doi:10.1109/gensips.2012.6507764.

Wu, Xiaogang, and Jake Y. Chen. 2012. “Molecular Interaction Networks: Topological and Functional Characterizations.” In Automation in Proteomics and Genomics, 145–74.

Wu, Xiaogang, Mohammad Al Hasan, and Jake Yue Chen. 2014. “Pathway and Network Analysis in Proteomics.” Journal of Theoretical Biology 362 (December): 44–52.

Wu, Xiaogang, Scott H. Harrison, and Jake Yue Chen. 2009. “Pattern Discovery in Breast Cancer Specific Protein Interaction Network.” Summit on Translational Bioinformatics 2009 (March): 1–5.

Wu, Xiaogang, Tianxiao Huan, Ragini Pandey, Tianshou Zhou, and Jake Y. Chen. 2009. “Finding Fractal Patterns in Molecular Interaction Networks: A Case Study in Alzheimer’s Disease.” International Journal of Computational Biology and Drug Design 2 (4): 340–52.

Zenklusen, J. C. 2014. “The Cancer Genome Atlas (TCGA): A Primer on Accessing the Data.” AACR Education Book 2014 (1): 155–60.

Zhang, Changsheng, and Luhua Lai. 2011. “SDOCK: A Global Protein-Protein Docking Program Using Stepwise Force-Field Potentials.” Journal of Computational Chemistry 32 (12): 2598–2612.

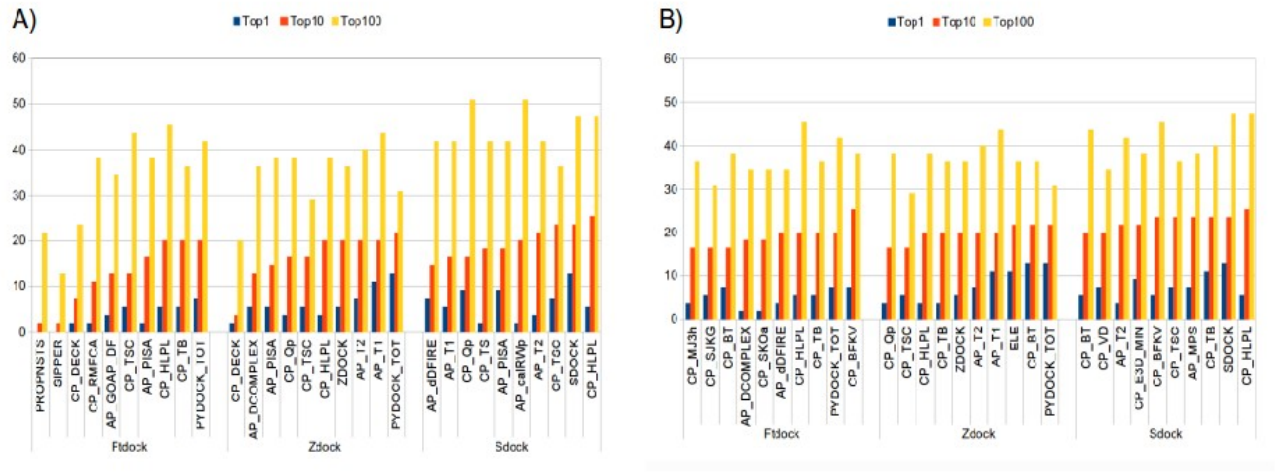
Zhang, Yang,. 2008. “I-TASSER Server for Protein 3D Structure Prediction.” BMC Bioinformatics 9 (1): 40.

Zhang, Yang. 2014. “Interplay of I-TASSER and QUARK for Template-Based and Ab Initio Protein Structure Prediction in CASP10.” *Proteins* 82 Suppl 2 (February): 175–87.

“I, myself, have killed six people. All random, all undetected, no way to trace them to me. And, let me tell you, there's nothin' like it. It's a great feeling. Yeah, I know, you're thinking. 'Aw, he's a comedian. He's just sayin' that stuff.' Good. That's exactly what I want you to think.”

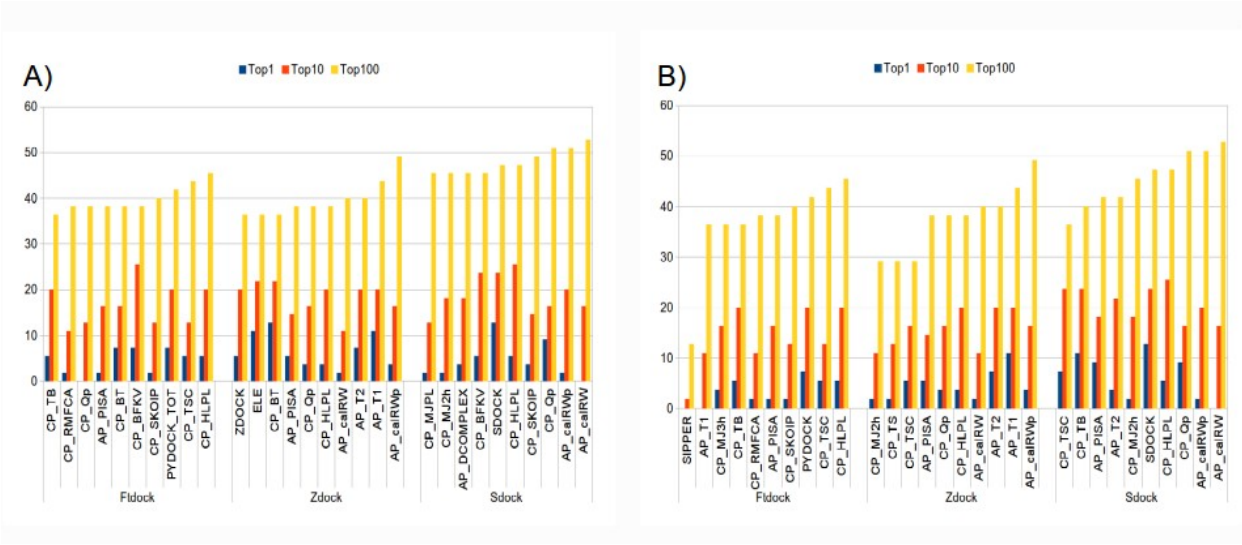
— George Carlin, Brain Droppings

# Chapter 8 Supplementary material



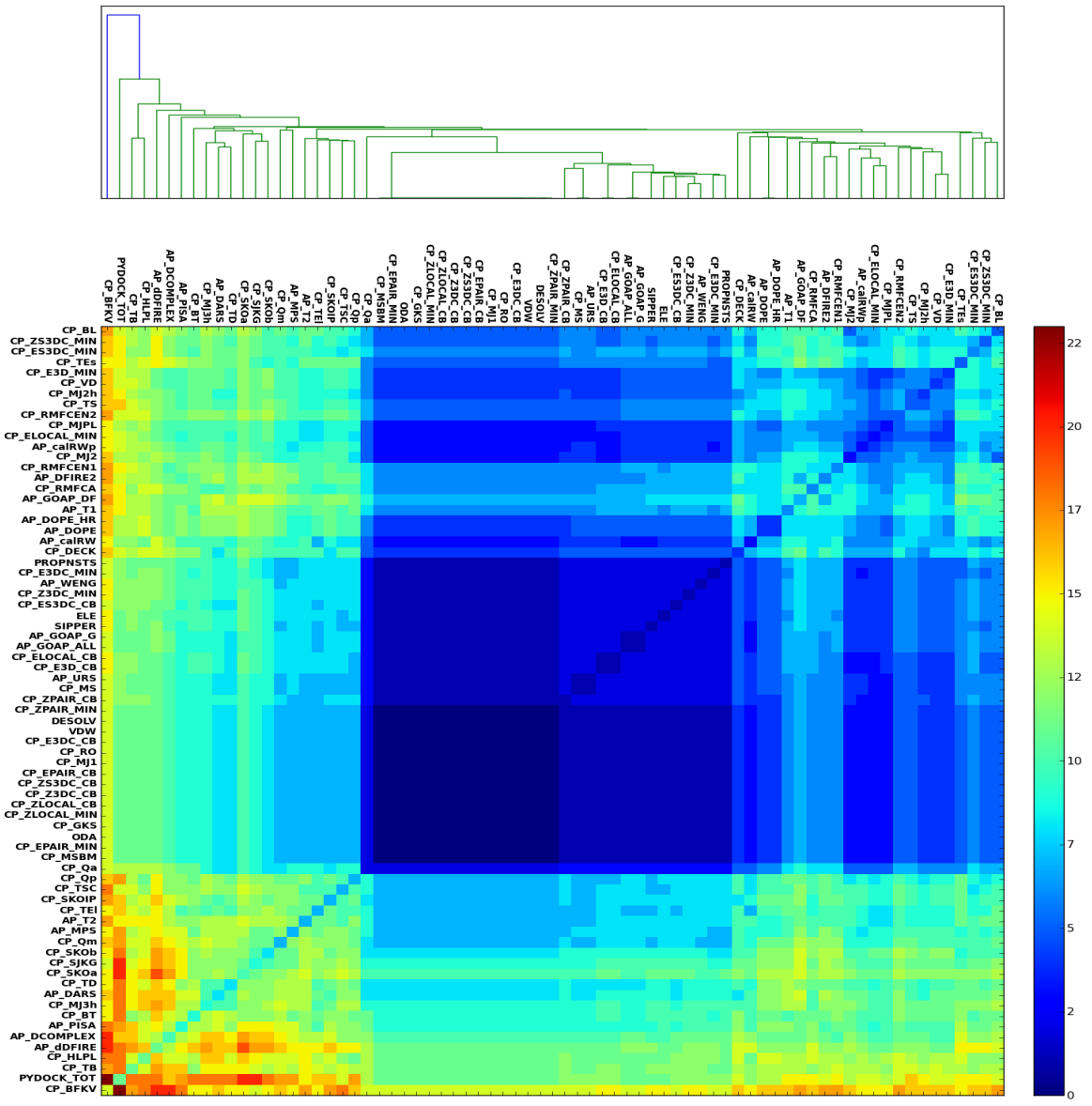

*Supplementary Figure 1:* **Comparison of the success rate of the Scoring Functions in the analysis in the BM 5.0**.

A) We used the BM 5.0 update as an external set of protein complexes to test possible overtraining of the best Scoring Functions found for the BM 4.0. B) The best scoring functions for the BM 5.0 are mainly coarse-grain Scoring Functions. Some of this new set of Scoring Functions are shared at least between two of the methods, like PYDOCK_TOT or CP_BFKV, and there are two Scoring Functions shared among all methods CP_HLPL and CP_TB . This latter showed a big success rate for the top100 ranking in the scorers set and CP_HLPL appear again shared in all methods, this indicates the robustness of both by the good scoring in the different set of decoys.

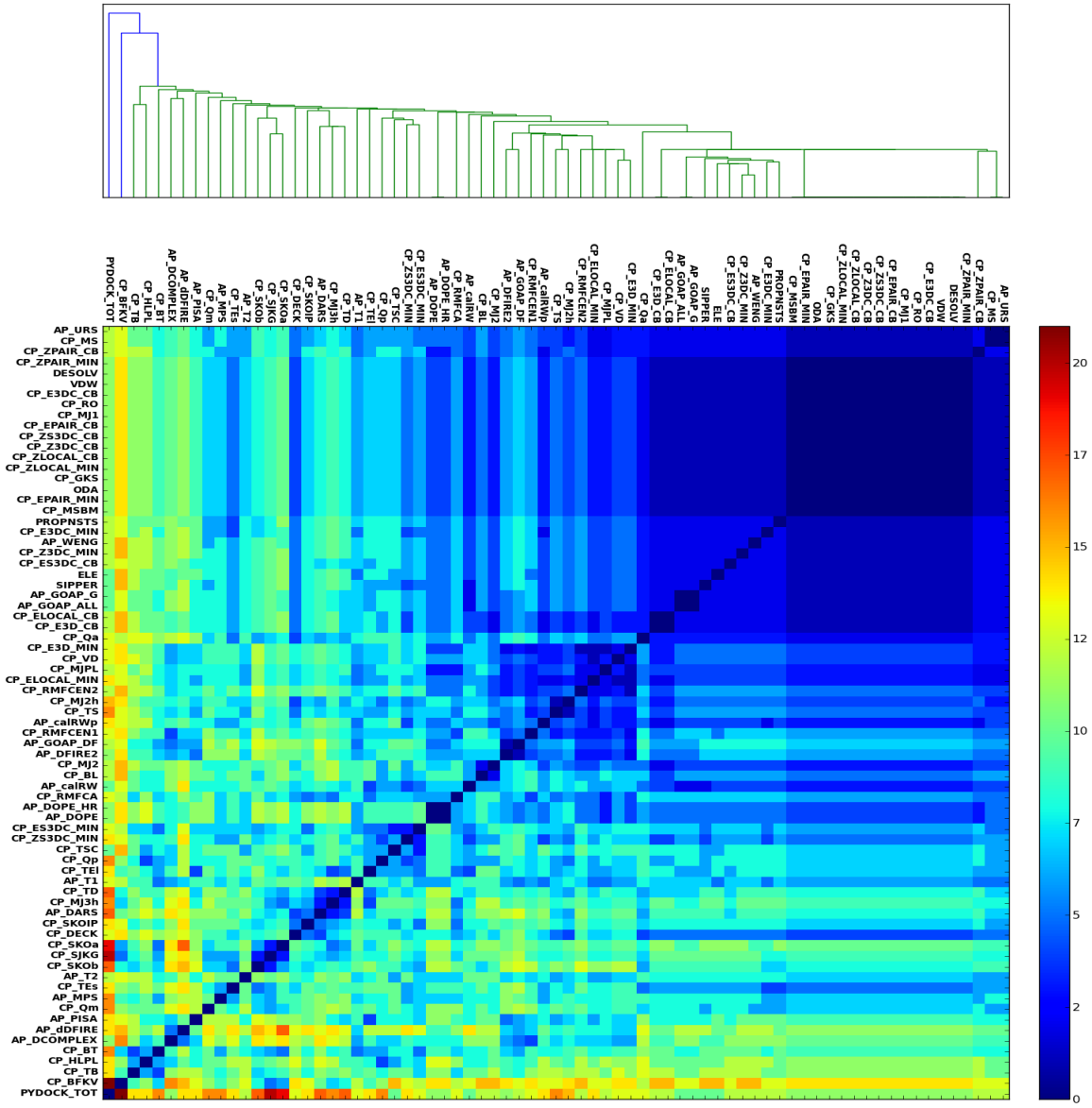



*Supplementary Figure 2* ***Comparison of the success rate in top100 ranking of the scoring functions in BM 5.0.***

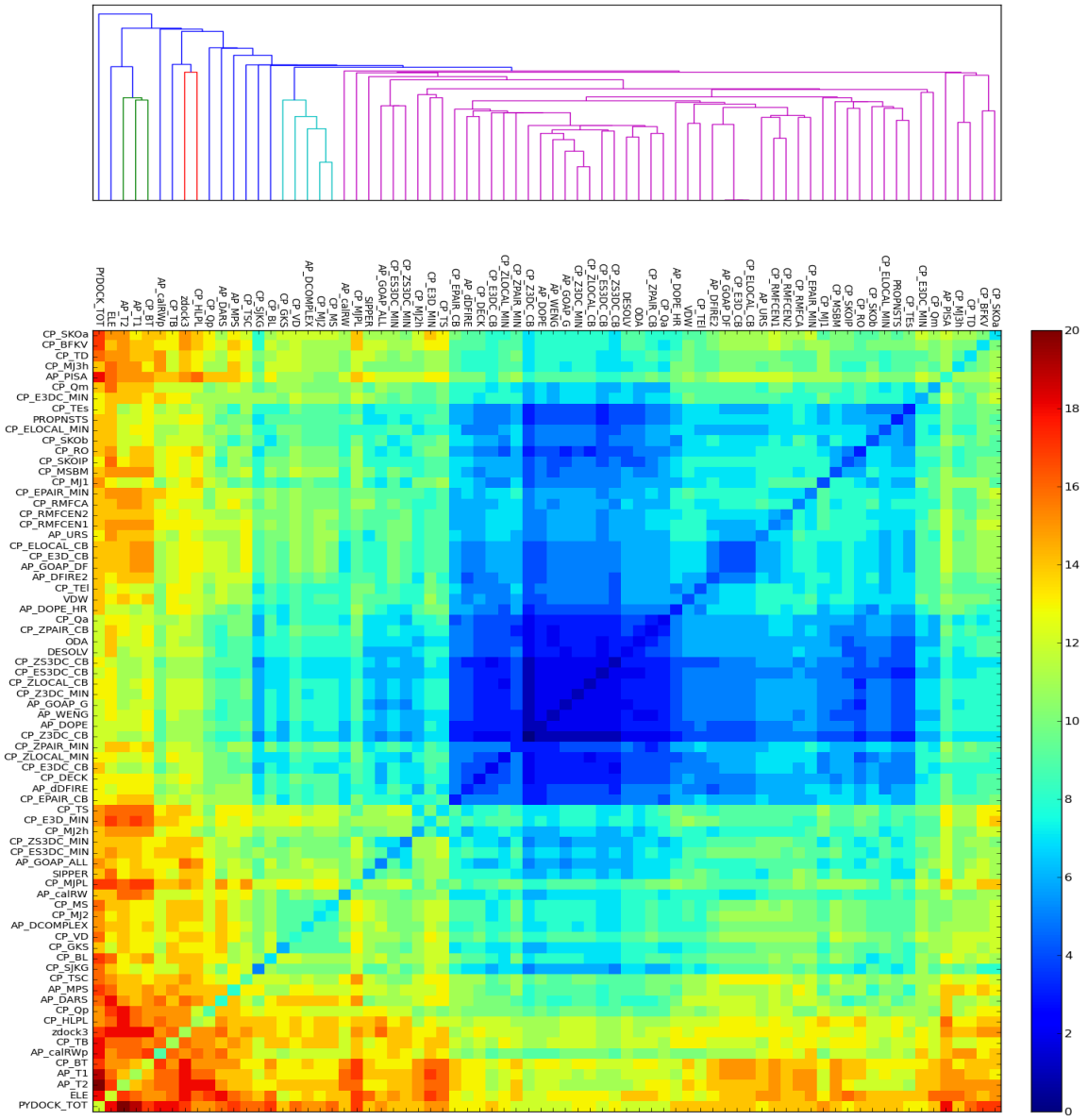
(A) In the BM 4.0, we found three Scoring Function: CP_TB, AP_PISA CP_HLPL have good success rate in top10 ranking and are shared among methods. Also, in general the atomistic Scoring Functions have better success rate than the coarse-grain Scoring Functions except CP_TSC. (B) The analysis in the BM 5.0 update changes the order of the ranking Scoring Functions. The best Scoring Functions in ZDOCK and SDOCK is AP_calRW and its refined version AP_calRWp, both showed big ranking power in the scorer set. With FTDock the best atomistic Scoring Functions is PYDOCK_TOT. The only Scoring Functions shared for all the three methods is CP_HLPL, this suggests this Scoring Functions is resilient to the possible changes caused by the different sampling methodologies.



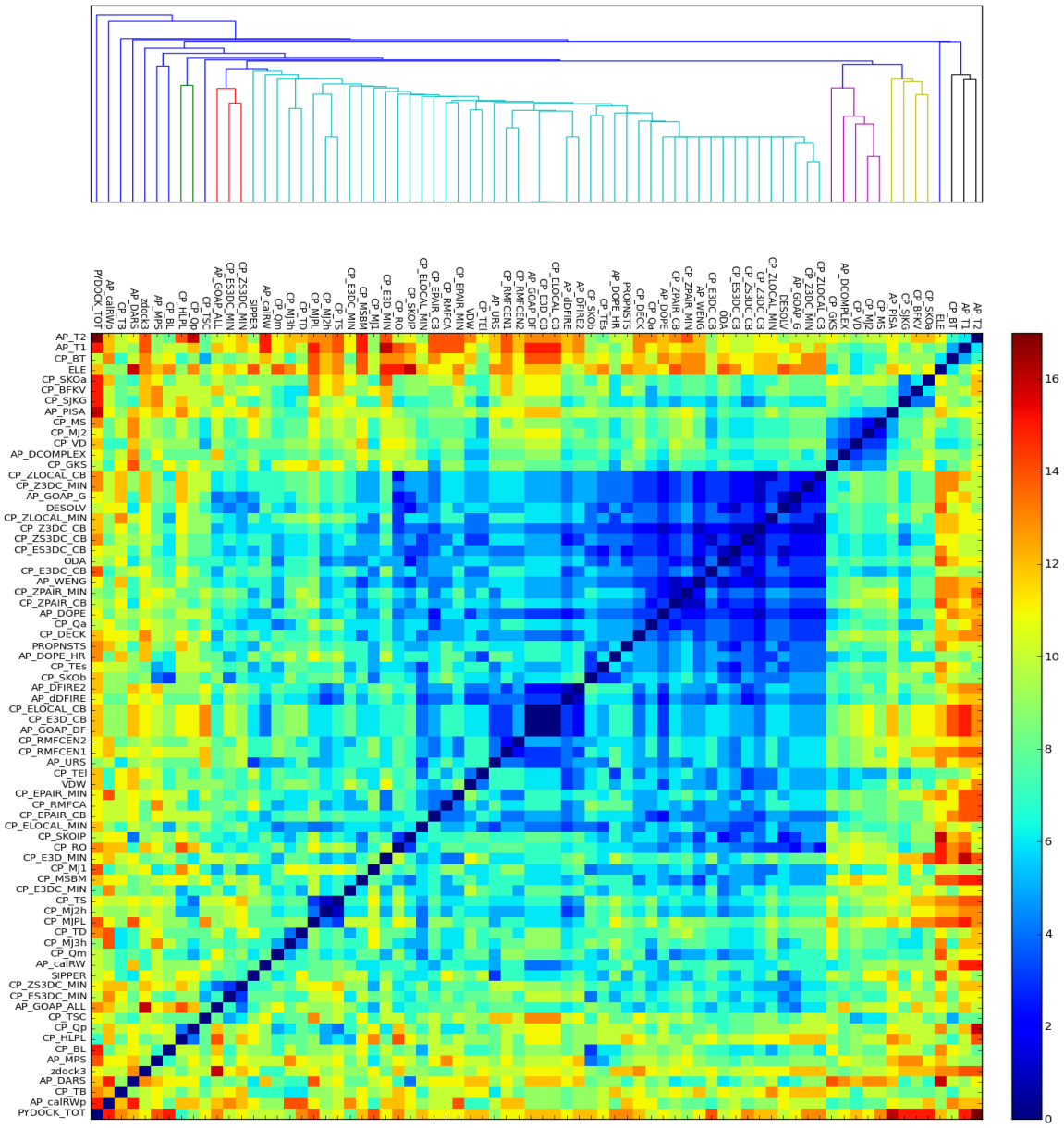

*Supplementary Figure 3* ***Union cardinalities heatmap of FTDock showing the relation between all the pairs of scoring functions***



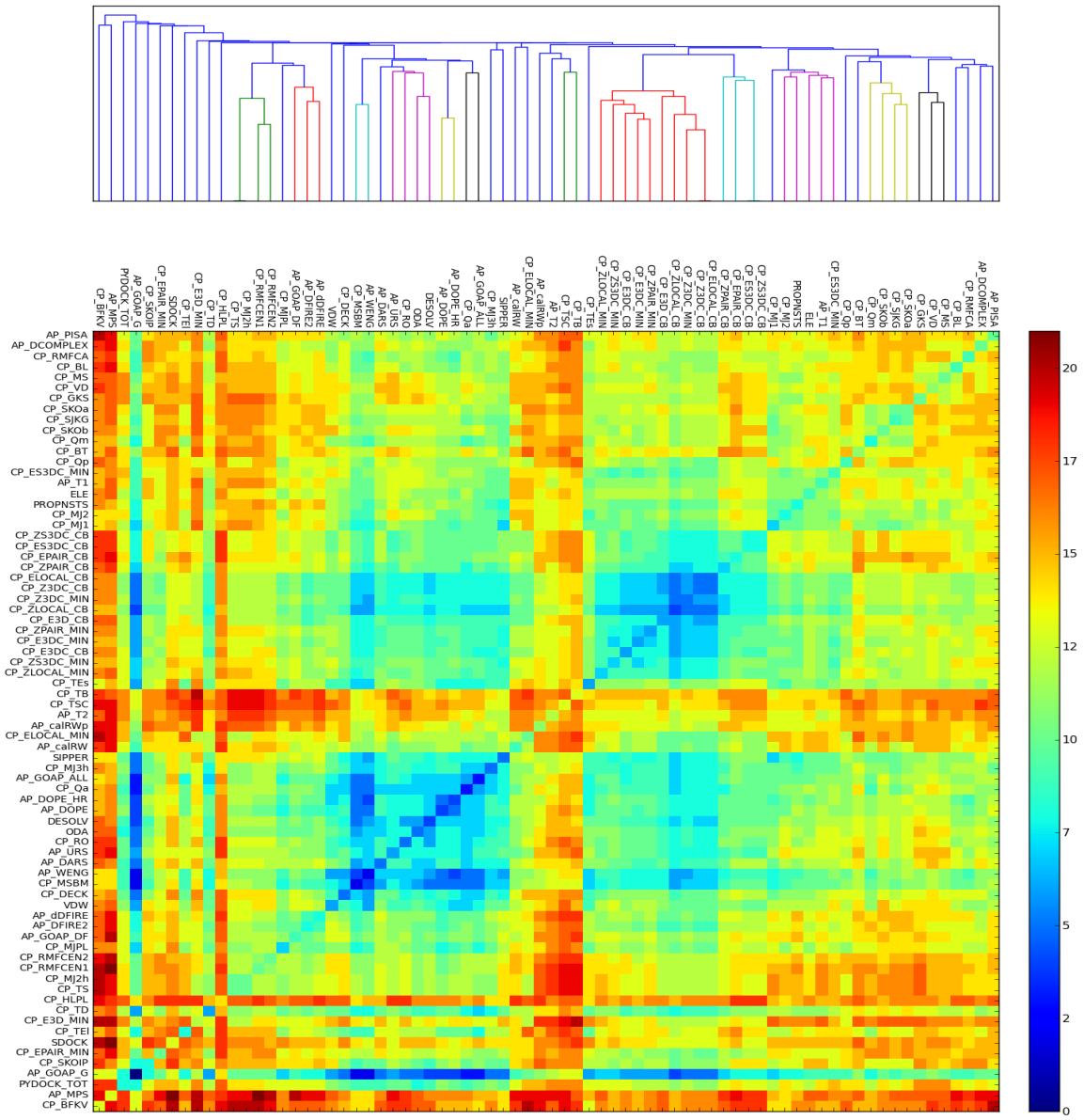

*Supplementary Figure 4* ***Symmetric difference cardinalities heatmap of FTDock showing the relation between all the pairs of scoring functions***



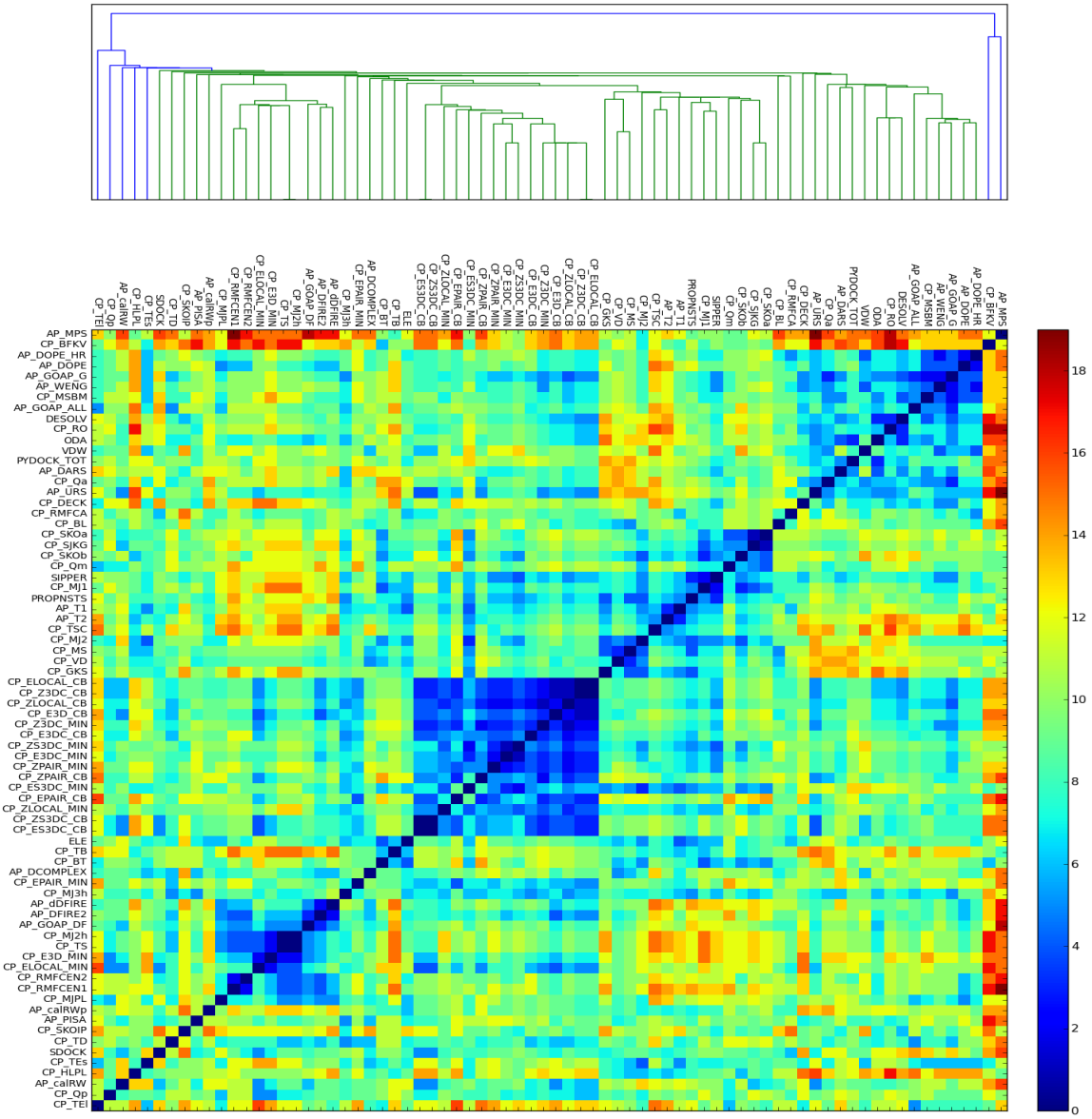

*Supplementary Figure 5* ***Union cardinalities heatmap of ZDOCK showing the relation between all the pairs of scoring functions***




*Supplementary Figure 6* ***Symetric difference cardinalities heatmap of ZDOCK showing the relation between all the pairs of scoring functions***




*Supplementary Figure 7* ***Union cardinalities heatmap of SDOCK showing the relation between all the pairs of scoring functions***




*Supplementary Figure 8* ***Symmetric difference cardinalities heatmap of SDOCK showing the relation between all the pairs of scoring functions***

*Supplementary Table 1* ***Docking success rates in top 1, top 10 and top 100, for the four docking pipelines before and after re-ranking, and a comparison to other docking protocols.***

*a) Re-ranked using the presented method, BM5 results are those when the BM5 complexes are used as external validation set, and BM4 results are cross-validation scores. b) Results from this study, prior to re-ranking using the same decoy set. c) Results from Vreven et al. d) Results fromTorchala et al. e) Results from Schneidman-Duhovny et al. f) Results from Chowdhury et al. . g)Results from Ohue et al. . h) Results from Huang, where a slightly different definition of near-native is used.*

| | | BM5 update (%) n=55 | | | BM4 (%) n=176 | | |
|---|---|---|---|---|---|---|---|
| Method | Reference | T1 | T10 | T100 | T1 | T10 | T100 |
| SwarmDock+re-rank | a | 24 | 45 | 69 | 30 | 49 | 65 |
| SwarmDock | b-c-d | 16 | 38 | 67 | 10 | 36 | 65 |
| pyDock+re-rank | a | 18 | 40 | 54 | 14 | 30 | 52 |
| pyDock | b | 7 | 20 | 42 | 6 | 18 | 40 |
| ZDOCK 3.0.1+re-rank | a | 20 | 27 | 45 | 20 | 38 | 65 |
| ZDOCK 3.0.1 | b | 5 | 20 | 36 | 13 | 26 | 47 |
| SDOCK+re-rank | a | 9 | 29 | 56 | 20 | 40 | 59 |
| SDOCK | b | 13 | 24 | 47 | 13 | 26 | 49 |
| ZDOCK3.0.2+IFACE | c | 5 | 27 | 53 | - | - | - |
| pyDock | c | 7 | 20 | 42 | - | - | - |
| HADDOCK | c | 9 | 20 | 40 | - | - | - |
| PatchDock+FireDock | e | - | - | - | 10 | 24 | 49 |
| F2Dock | f | - | - | - | 13 | 25 | 38 |
| ZDOCK 3.0.2 | f | - | - | - | 7 | 22 | 42 |
| MegaDock 1.0 | g | - | - | - | 0 | 2 | 6 |
| MegaDock 2.0 | g | - | - | - | 1 | 3 | 10 |
| MegaDock 2.1 | g | - | - | - | 1 | 5 | 13 |
| ZDOCK 3.0 | g | - | - | - | 7 | 13 | 27 |
| ZDOCK 3.0.2 | h | - | - | - | 12 | 31 | 52 |
| SDOCK | h | - | - | - | 10 | 23 | 46 |
| PIPER | h | - | - | - | 9 | 21 | 40 |
| FRODOCK | h | - | - | - | 5 | 19 | 46 |
| ATTRACT:LJ | h | - | - | - | 5 | 19 | 47 |
| ATTRACT | h | - | - | - | 5 | 18 | 43 |
| ZDOCK 1.3 | h | - | - | - | 7 | 15 | 41 |
| ZDOCK 2.3.2 | h | - | - | - | 6 | 14 | 38 |
| HEX | h | - | - | - | 4 | 11 | 25 |
| DOT | h | - | - | - | 2 | 10 | 27 |
| PatchDock | h | - | - | - | 3 | 7 | 23 |
| MolFit/GH | h | - | - | - | 2 | 7 | 25 |
| ZDOCK 2.1 | h | - | - | - | 1 | 7 | 20 |
| HEX/G | h | - | - | - | 0 | 4 | 16 |
| MolFit/G | h | - | - | - | 1 | 3 | 19 |
| GRAMM | h | - | - | - | 0 | 3 | 11 |
| FTDock/G | h | - | - | - | 1 | 2 | 11 |
| FTDock | h | - | - | - | 1 | 2 | 11 |

*Supplementary Table 2:* ***Scoring functions use from the Ccharppi server and their reference.***

| Scoring Function | Description | Reference |
|---|---|---|
| CP_BFKV | Contact potential calculated between intermolecular residues | Proteins 59(1):49 (2005) and BMC Bioinformatics 11:92 (2010) |
| CP_BL | Contact potential calculated between intermolecular residues | Proteins 59(1):49 (2005) and BMC Bioinformatics 11:92 (2010) |
| CP_BT | Contact potential calculated between intermolecular residues | Proteins 59(1):49 (2005) and BMC Bioinformatics 11:92 (2010) |
| CP_GKS | Contact potential calculated between intermolecular residues | Proteins 59(1):49 (2005) and BMC Bioinformatics 11:92 (2010) |
| CP_HLPL | Contact potential calculated between intermolecular residues | Proteins 59(1):49 (2005) and BMC Bioinformatics 11:92 (2010) |
| CP_MJPL | Contact potential calculated between intermolecular residues | Proteins 59(1):49 (2005) and BMC Bioinformatics 11:92 (2010) |
| CP_MJ3h | Contact potential calculated between intermolecular residues | Proteins 59(1):49 (2005) and BMC Bioinformatics 11:92 (2010) |
| CP_MJ2h | Contact potential calculated between intermolecular residues | Proteins 59(1):49 (2005) and BMC Bioinformatics 11:92 (2010) |
| CP_MJ1 | Contact potential calculated between intermolecular residues | Proteins 59(1):49 (2005) and BMC Bioinformatics 11:92 (2010) |
| CP_MJ2 | Contact potential calculated between intermolecular residues | Proteins 59(1):49 (2005) and BMC Bioinformatics 11:92 (2010) |
| CP_MSBM | Contact potential calculated between intermolecular residues | Proteins 59(1):49 (2005) and BMC Bioinformatics 11:92 (2010) |
| CP_MS | Contact potential calculated between intermolecular residues | Proteins 59(1):49 (2005) and BMC Bioinformatics 11:92 (2010) |
| CP_Qa | Contact potential calculated between intermolecular residues | Proteins 59(1):49 (2005) and BMC Bioinformatics 11:92 (2010) |
| CP_Qm | Contact potential calculated between intermolecular residues | Proteins 59(1):49 (2005) and BMC Bioinformatics 11:92 (2010) |
| CP_Qp | Contact potential calculated between intermolecular residues | Proteins 59(1):49 (2005) and BMC Bioinformatics 11:92 (2010) |
| CP_RO | Contact potential calculated between intermolecular residues | Proteins 59(1):49 (2005) and BMC Bioinformatics 11:92 (2010) |
| CP_SKOb | Contact potential calculated between | Proteins 59(1):49 (2005) and BMC |

| | | |
|---|---|---|
| | intermolecular residues | Bioinformatics 11:92 (2010) |
| CP_SKOa | Contact potential calculated between intermolecular residues | Proteins 59(1):49 (2005) and BMC Bioinformatics 11:92 (2010) |
| CP_SJKG | Contact potential calculated between intermolecular residues | Proteins 59(1):49 (2005) and BMC Bioinformatics 11:92 (2010) |
| CP_TD | Contact potential calculated between intermolecular residues | Proteins 59(1):49 (2005) and BMC Bioinformatics 11:92 (2010) |
| CP_TEl | Contact potential calculated between intermolecular residues | Proteins 59(1):49 (2005) and BMC Bioinformatics 11:92 (2010) |
| CP_TEs | Contact potential calculated between intermolecular residues | Proteins 59(1):49 (2005) and BMC Bioinformatics 11:92 (2010) |
| CP_TS | Contact potential calculated between intermolecular residues | Proteins 59(1):49 (2005) and BMC Bioinformatics 11:92 (2010) |
| CP_VD | Contact potential calculated between intermolecular residues | Proteins 59(1):49 (2005) and BMC Bioinformatics 11:92 (2010) |
| CP_TSC | The residue level interaction two-step potential | BMC Struct biol 10:40 (2010). |
| CP_SKOIP | The residue level interaction contact potential | Biophys. J. 84(3):1895 (2003). |
| AP_DCOMPLEX | The DComplex potential | Proteins 56:93 (2004). |
| AP_dDFIRE | Interaction energy calculated using the dDFIRE potential | Proteins 72:793 (2008). |
| AP_DFIRE2 | Interaction energy calculated using the DFIRE2 potential | Protein Science 17:1212 (2008). |
| CP_RMFCEN1 | The 6bin-HRSC centroid-centroid potential | Proteins 70(3):950 (2006). |
| CP_RMFCEN2 | The 7bin-HRSC centroid-centroid potential | Proteins 70(3):950 (2006). |
| CP_RMFCA | The C_alpha-C_alpha potential | Proteins 65(3):726 (2006) |
| CP_TB | The residue level interaction contact potential | Proteins 62(4):970 (2006). |
| CP_TSC | The residue level interaction two-step potential | BMC Struct biol 10:40 (2010). |
| AP_T1 | The first atomic two-step potential | BMC Struct biol 10:40 (2010). |
| AP_T2 | The second atomic two-step potential | BMC Struct biol 10:40 (2010). |

| | | |
|---|---|---|
| AP_DOPE | The DOPE statistical potential | Protein Sci. 15(11):2507 (2006). |
| ELE | Total electrostatic energy as calculated using PyDock | Proteins 68:503 (2007) and Protein 69:852 (2007) |
| DESOLV | Desolvation energy as calculated using PyDock | Proteins 68:503 (2007) and Protein 69:852 (2007) |
| VDW | Van der Waals energy as calculated using PyDock | Proteins 68:503 (2007) and Protein 69:852 (2007) |
| PYDOCK_TOT | Total pyDock energy | Proteins 68:503 (2007) and Protein 69:852 (2007) |
| ODA | The optimal docking area (ODA) score | Int. J. Data Mining Bioinf. 3:55 (2009) and J. Chem. Inf. Mod. 51:370 (2011). |
| PROPNSTS | Amino acid propensity score | J. Chem. Inf. Mod. 51:370 (2011). |
| SIPPER | The SIPPER potential | J. Chem. Inf. Mod. 51:370 (2011). |
| AP_DARS | The DARS potential | Biophys J. 2008 95(9):4217-27 |
| AP_URS | The URS potential | Biophys J. 2008 95(9):4217-26 |
| AP_MPS | The MPS potential | Biophys J. 2008 95(9):4217-25 |
| AP_WENG | The pair-wise statistical potential implemented in Zdock | Proteins 2007 69(3):511-20. |
| CP_DECK | The residue level distance-dependent potential | BMC Bioinformatics. 2011 12:280. |
| CP_ZPAIR_CB | The E_pair Z-score C_beta potential | Protein Sci. 2011 20(3):529-41. |
| CP_ZLOCAL_CB | The E_local Z-score C_beta potential | Protein Sci. 2011 20(3):529-41. |
| CP_ZS3DC_CB | The E_ZS3DC z-score C_beta potential | Protein Sci. 2011 20(3):529-41. |
| CP_Z3DC_CB | The E_3DC Z-score C_beta potential | Protein Sci. 2011 20(3):529-41. |
| CP_EPAIR_CB | The E_pair C_beta potential | Protein Sci. 2011 20(3):529-41. |
| CP_ELOCAL_CB | The E_local C_beta potential | Protein Sci. 2011 20(3):529-41. |
| CP_ES3DC_CB | The E_ZS3DC C_beta potential | Protein Sci. 2011 20(3):529-41. |
| CP_E3DC_CB | The E_3DC C_beta potential | Protein Sci. 2011 20(3):529-41. |
| CP_E3D_CB | The E_3D C_beta potential | Protein Sci. 2011 20(3):529-41. |
| CP_ZPAIR_MIN | The E_pair Z-score R_min potential | Protein Sci. 2011 20(3):529-41. |
| CP_ZLOCAL_MIN | The E_local Z-score R_min potential | Protein Sci. 2011 20(3):529-41. |

| | | |
|---|---|---|
| CP_ZS3DC_MIN | The E_ZS3DC z-score R_min potential | Protein Sci. 2011 20(3):529-41. |
| CP_Z3DC_MIN | The E_3DC Z-score R_min potential | Protein Sci. 2011 20(3):529-41. |
| CP_EPAIR_MIN | The E_pair R_min potential | Protein Sci. 2011 20(3):529-41. |
| CP_ELOCAL_MIN | The E_local R_min potential | Protein Sci. 2011 20(3):529-41. |
| CP_ES3DC_MIN | The E_ZS3DC R_min potential | Protein Sci. 2011 20(3):529-41. |
| CP_E3DC_MIN | The E_3DC R_min potential | Protein Sci. 2011 20(3):529-41. |
| CP_E3D_MIN | The E_3D R_min potential | Protein Sci. 2011 20(3):529-41. |
| AP_calRW | The calRW distance-dependent atomic potential | PloS One. 2010 5(10):e15386. |
| AP_calRWp | The calRWplus orientation-dependent atomic potential | PloS One. 2010 5(10):e15386. |
| AP_GOAP_ALL | The total GOAP energy | Biophys J. 2011 101(8): 2043-2052. |
| AP_GOAP_DF | The DFIRE term in the GOAP energy | Biophys J. 2011 101(8): 2043-2052. |
| AP_GOAP_G | The GOAP_ag term in the GOAP energy | Biophys J. 2011 101(8): 2043-2052. |
| AP_PISA | The PISA score | Proteins 81(4):592 (2013). |

*Supplementary Table 3* ***Union cardinality of top 50 perfoming pairs of scoring functions when we combined zscores from the three different FFT methods ordered by union columun.***

| FTDock | ZDOCK | SDOCK | UNION |
|---|---|---|---|
| CP_HLPL and CP_DECK | PYDOCK_TOT and ZDOCK | CP_QP and AP_calRWp | 30 |
| PYDOCK_TOT and AP_GOAP_DF | PYDOCK_TOT and ZDOCK | AP_T2 and CP_TSC | 29 |
| AP_PISA and CP_TSC | PYDOCK_TOT and ZDOCK | CP_QP and AP_calRWp | 29 |
| PYDOCK_TOT and AP_GOAP_DF | PYDOCK_TOT and AP_T1 | AP_T2 and CP_TSC | 28 |
| PYDOCK_TOT and AP_GOAP_DF | CP_DECK and ZDOCK | AP_T2 and CP_TSC | 28 |
| PYDOCK_TOT and AP_GOAP_DF | AP_T2 and PYDOCK_TOT | AP_T2 and CP_TSC | 28 |
| PYDOCK_TOT and AP_GOAP_DF | AP_T2 and CP_DECK | AP_T2 and CP_TSC | 28 |
| PYDOCK_TOT and AP_GOAP_DF | AP_PISA and PYDOCK_TOT | AP_T2 and CP_TSC | 28 |
| PYDOCK_TOT and AP_GOAP_DF | AP_PISA and CP_TSC | AP_T2 and CP_TSC | 28 |
| CP_RMFCA and AP_GOAP_DF | PYDOCK_TOT and ZDOCK | AP_T2 and CP_TSC | 28 |
| CP_HLPL and PYDOCK_TOT | PYDOCK_TOT and ZDOCK | CP_QP and AP_calRWp | 28 |
| CP_HLPL and PYDOCK_TOT | PYDOCK_TOT and ZDOCK | AP_T2 and CP_TSC | 28 |
| CP_HLPL and PYDOCK_TOT | AP_PISA and CP_TSC | CP_QP and CP_TS | 28 |
| CP_HLPL and PYDOCK_TOT | AP_PISA and CP_TSC | AP_calRWp and CP_TS | 28 |
| CP_HLPL and CP_RMFCA | PYDOCK_TOT and ZDOCK | AP_T2 and CP_TSC | 28 |
| CP_HLPL and CP_DECK | PYDOCK_TOT and ZDOCK | CP_QP and CP_TS | 28 |
| CP_HLPL and CP_DECK | PYDOCK_TOT and ZDOCK | CP_QP and CP_HLPL | 28 |
| CP_HLPL and CP_DECK | PYDOCK_TOT and ZDOCK | CP_QP and AP_PISA | 28 |
| CP_HLPL and CP_DECK | PYDOCK_TOT and ZDOCK | CP_QP and AP_dDFIRE | 28 |
| CP_HLPL and CP_DECK | PYDOCK_TOT and ZDOCK | CP_HLPL and AP_PISA | 28 |
| CP_HLPL and CP_DECK | PYDOCK_TOT and ZDOCK | AP_T2 and CP_TSC | 28 |
| CP_HLPL and CP_DECK | PYDOCK_TOT and ZDOCK | AP_dDFIRE and CP_HLPL | 28 |
| CP_HLPL and CP_DECK | PYDOCK_TOT and ZDOCK | AP_calRWp and CP_TSC | 28 |
| CP_HLPL and CP_DECK | PYDOCK_TOT and ZDOCK | AP_calRWp and CP_TS | 28 |
| CP_HLPL and CP_DECK | PYDOCK_TOT and ZDOCK | AP_calRWp and AP_dDFIRE | 28 |
| CP_HLPL and CP_DECK | AP_T2 and PYDOCK_TOT | CP_QP and AP_calRWp | 28 |
| CP_HLPL and CP_DECK | AP_PISA and CP_TSC | CP_QP and CP_TS | 28 |
| CP_HLPL and CP_DECK | AP_PISA and CP_TSC | CP_QP and AP_PISA | 28 |
| CP_HLPL and CP_DECK | AP_PISA and CP_TSC | CP_QP and AP_calRWp | 28 |
| CP_HLPL and CP_DECK | AP_PISA and CP_TSC | CP_HLPL and AP_PISA | 28 |
| CP_HLPL and CP_DECK | AP_PISA and CP_TSC | AP_calRWp and CP_TS | 28 |
| CP_HLPL and AP_PISA | PYDOCK_TOT and ZDOCK | CP_QP and AP_calRWp | 28 |
| CP_HLPL and AP_PISA | PYDOCK_TOT and ZDOCK | AP_T2 and CP_TSC | 28 |
| CP_HLPL and AP_PISA | PYDOCK_TOT and ZDOCK | AP_calRWp and CP_TSC | 28 |
| CP_HLPL and AP_PISA | AP_PISA and CP_TSC | CP_QP and AP_PISA | 28 |
| CP_HLPL and AP_PISA | AP_PISA and CP_TSC | CP_QP and AP_calRWp | 28 |
| CP_HLPL and AP_PISA | AP_PISA and CP_TSC | CP_HLPL and AP_PISA | 28 |
| CP_HLPL and AP_PISA | AP_PISA and CP_TSC | AP_T2 and CP_TSC | 28 |
| CP_HLPL and AP_GOAP_DF | PYDOCK_TOT and ZDOCK | AP_T2 and CP_TSC | 28 |
| AP_PISA and PYDOCK_TOT | PYDOCK_TOT and ZDOCK | CP_QP and AP_calRWp | 28 |
| AP_PISA and PYDOCK_TOT | PYDOCK_TOT and ZDOCK | AP_calRWp and CP_HLPL | 28 |
| AP_PISA and CP_RMFCA | PYDOCK_TOT and ZDOCK | AP_T2 and CP_TSC | 28 |
| AP_PISA and AP_GOAP_DF | PYDOCK_TOT and ZDOCK | AP_T2 and CP_TSC | 28 |
| SIPPER and CP_TB | PYDOCK_TOT and ZDOCK | CP_QP and CP_HLPL | 27 |
| SIPPER and CP_TB | PYDOCK_TOT and ZDOCK | CP_HLPL and AP_PISA | 27 |
| SIPPER and AP_GOAP_DF | PYDOCK_TOT and ZDOCK | CP_QP and CP_HLPL | 27 |
| SIPPER and AP_GOAP_DF | PYDOCK_TOT and ZDOCK | CP_QP and AP_PISA | 27 |
| SIPPER and AP_GOAP_DF | PYDOCK_TOT and ZDOCK | CP_QP and AP_calRWp | 27 |
| SIPPER and AP_GOAP_DF | PYDOCK_TOT and ZDOCK | CP_HLPL and AP_PISA | 27 |

*Supplementary Table 4:* ***Union Cardinality of the tripplets used to re score BM 5.0 update***

| FTDock | ZDOCK | SDOCK | UNION |
|---|---|---|---|
| PYDOCK_TOT and CP_TB | PYDOCK_TOT and ZDOCK | SDOCK and CP_TSC | 26 |
| CP_HLPL and PYDOCK_TOT | AP_T2 and CP_DECK | AP_calRWp and SDOCK | 26 |
| CP_HLPL and PYDOCK_TOT | AP_PISA and AP_T1 | SDOCK and AP_PISA | 25 |
| PYDOCK_TOT and CP_TB | AP_T2 and PYDOCK_TOT | SDOCK and CP_TSC | 25 |
| CP_HLPL and CP_DECK | PYDOCK_TOT and CP_TSC | SDOCK and AP_PISA | 25 |
| PYDOCK_TOT and AP_GOAP_DF | AP_PISA and ZDOCK | SDOCK and AP_PISA | 25 |
| PYDOCK_TOT and AP_GOAP_DF | AP_PISA and PYDOCK_TOT | SDOCK and AP_PISA | 25 |
| PYDOCK_TOT and AP_GOAP_DF | AP_T1 and ZDOCK | SDOCK and AP_PISA | 25 |
| CP_HLPL and PYDOCK_TOT | PYDOCK_TOT and AP_DCOMPLEX | AP_calRWp and SDOCK | 25 |
| CP_HLPL and PYDOCK_TOT | AP_T2 and ZDOCK | AP_calRWp and SDOCK | 25 |
| CP_HLPL and PYDOCK_TOT | AP_T2 and AP_PISA | AP_calRWp and SDOCK | 25 |
| PYDOCK_TOT and CP_TB | AP_PISA and CP_TSC | SDOCK and AP_T1 | 24 |
| AP_PISA and CP_TB | PYDOCK_TOT and CP_TSC | SDOCK and AP_PISA | 24 |
| PYDOCK_TOT and CP_TB | AP_PISA and ZDOCK | SDOCK and CP_TSC | 24 |
| PYDOCK_TOT and CP_TB | PYDOCK_TOT and CP_TSC | CP_Qp and SDOCK | 24 |
| PYDOCK_TOT and AP_GOAP_DF | AP_T2 and CP_TSC | SDOCK and AP_PISA | 24 |
| PYDOCK_TOT and AP_GOAP_DF | PYDOCK_TOT and AP_DCOMPLEX | SDOCK and AP_PISA | 24 |
| PYDOCK_TOT and CP_TB | CP_Qp and PYDOCK_TOT | SDOCK and CP_TSC | 24 |
| CP_HLPL and PYDOCK_TOT | AP_DCOMPLEX_AP_T1 | AP_calRWp and SDOCK | 24 |
| PYDOCK_TOT and CP_TB | AP_PISA and ZDOCK | CP_Qp and SDOCK | 23 |
| PYDOCK_TOT and CP_TB | AP_PISA and AP_T1 | CP_Qp and SDOCK | 23 |
| PYDOCK_TOT and AP_GOAP_DF | CP_Qp and AP_DCOMPLEX | SDOCK and AP_PISA | 23 |
| PYDOCK_TOT and CP_TB | PYDOCK_TOT and CP_TSC | SDOCK and CP_HLPL | 23 |
| PYDOCK_TOT and CP_TB | PYDOCK_TOT and CP_TSC | SDOCK and CP_TSC | 23 |
| PYDOCK_TOT and AP_GOAP_DF | AP_T2 and AP_DCOMPLEX | SDOCK and AP_PISA | 23 |
| PYDOCK_TOT and CP_TB | CP_Qp and PYDOCK_TOT | AP_T2 and SDOCK | 23 |
| PYDOCK_TOT and AP_GOAP_DF | CP_Qp and PYDOCK_TOT | SDOCK and CP_HLPL | 22 |
| PYDOCK_TOT and CP_TB | CP_Qp and AP_T2 | AP_T2 and SDOCK | 22 |
| PYDOCK_TOT and CP_TB | AP_T2 and AP_DCOMPLEX | SDOCK and CP_TSC | 22 |
| PYDOCK_TOT and CP_TB | AP_DCOMPLEX_AP_T1 | SDOCK and CP_TSC | 22 |

“You must unlearn what you have learned.”

– Yoda

# Chapter 9 Thesis advisor report

The present PhD thesis by Didier Barradas Bautista produced four academic articles, three as first author and one as second author. Three of the articles are submitted to journal with impact factor from 2.499 to 5.766 as indexed in ISI. The remaining article is soon to be submitted. The work has been presented to the scientific community in local and international congresses with talks or poster.

**•Research articules**

2016 **A large scale characterization of disease-related variants in the structural human interactome using high-throughput docking calculations .**Barradas-Bautista D and Fernández-Recio J. **In preparation**

2016 **A systematic analysis of scoring functions in rigid-body docking: the delicate balance between the predictive rate improvement and the risk of overtraining.** Barradas-Bautista D, Moal I, Fernández-Recio J. **Submitted**

2016 **Structural modeling of protein-protein interfaces for the functional characterization of disease-related amino acid mutations .**Barradas-Bautista D and Fernández-Recio J. **Submitted**

2016 **Web-search based integration of biophysical models for protein assembly selection**. Moal I, Barradas-Bautista D,Jiménez-García B, Torchala M, Van der Velde A, Vreven T, Weng Z, Bates PA and Fernández-Recio J. **Submited**

**•Congress and workshops attendance**

2016 POSTER: **FROM NETWORKS TO INTERACTIONS : 3D DISEASOMES AND HOTSPOT PREDICTIONS TO IDENTIFY EDGETIC MUTATIONS.** BIOINTERACTOMICS - FEBS IUBMB WORKSHOP **,SPAIN**

2016 POSTER: **PREDICTING EDGETIC MUTATIONS IN STRUCTURAL DISEASOMES USING PROTEIN-PROTEIN DOCKING SIMULATIONS: FROM NETWORK TO 3D INTERACTIONS AND BACK**. XVII INTERNATIONAL CONGRESS OF SYSTEM BIOLOGY, **SPAIN**

2016 TALK: **A STRATEGY TO MIX DIFFERENT BIOPHYSICAL SCORING FUNCTIONS FUSING THE RANKING POWER OF FFT-BASED PROTEIN-PROTEIN DOCKING PROTOCOLS.** 5TH INTERNATIONAL IBERIAN BIOPHYSICS CONGRESS**, PORTUGAL**

2016 TALK: **STRUCTURAL DISEASOMES AND HOT-SPOT PREDICTION ENABLE DETECTION OF NETWORK-ATTACKING MUTATIONS** XIII SYMPOSIUM ON BIOINFORMATICS, **SPAIN**

2016 TALK: **DOCKING THROUGH DEMOCRACY RE-RANKING PROTEIN-PROTEIN DECOYS WITH A VOTING SYSTEM.** 3RD BSC INTERNATIONAL DOCTORAL SYMPOSIUM, **SPAIN**

2016 POSTER: **NETWORK-ATTACKING MUTATIONS DETECTED IN STRUCTURAL DISEASOMES USING HOT-SPOT PREDICTION.** 1ST EUROPEAN CONFERENCE-ON-TRANSLATIONAL BIOINFORMATICS, **DENMARK**

2016 POSTER: **MULTIVARIATE BIOPHYSICAL COMBINATIONS TO ENHANCE RIGID BODY PROTEIN-PROTEIN DOCKING**. INTERNATIONAL CONFERENCE ON MOLECULAR RECOGNITION **SPAIN**

2015 TALK: **INTERFACE HOT SPOT PREDICTION TO DETECT MUTATIONS ALTERING THE PROTEIN-PROTEIN INTERACTIONS.** IX STRUCTURE AND FUNCTION OF

PROTEINS NETWORK MEETING, **SPAIN**

2015 POSTER: **A COMPREHENSIVE ANALYSIS OF SCORING FUNCTIONS FOR PROTEIN-PROTEIN DOCKING.** INTERNACIONAL, 29TH ANNUAL SYMPOSIUM OF TH PROTEIN SOCIETY, **SPAIN**

2015 POSTER: **HOT-SPOT INTERFACE PREDICTIONS TO IDENTIFY EDGETIC MUTATIONS**; INTERNACIONAL, XXXVIII CONGRESO DE LA SEBBM VALENCIA; SPAIN

2015 TALK: **PREDICTION OF INTERFACE HOT-SPOT RESIDUES TO CHARACTERIZE PATHOLOGICAL MUTATIONS IN PROTEIN- PROTEIN INTERACTIONS** .XXII JORNADES DE BIOLOGIA MOLECULAR ; SOCIETAT CATALANA DE BIOLOGIA, **SPAIN**

2015 TALK : **ASSESSMENT OF SCORING FUNCTIONS PERFORMANCE TO RE-RANK DOCKING DECOYS FROM FFT(FAST FOURIER TRANSFORM) PROGRAMS**,2ND BSC INTERNATIONAL DOCTORAL SYMPOSIUM, **SPAIN.**

2014 POSTER: **SCORING DOCKING CONFORMATIONS USING STRUCTURAL ALIGNMENT OF PROTEIN-PROTEIN INTERFACES.** INTERNACIONAL SERGIO MARES-SAMANO, DIDIER BARRADAS-BAUTISTA; JUAN FERNANDEZ-RECIO , XII JORNADAS DE BIOINFORMATICA, **SPAIN.**