# *In Silico* Prediction of Modular Domain-Peptide Interactions

**Dissertation**
zur Erlangung des akademischen Grades
doctor rerum naturalium (Dr. rer. nat.)

vorgelegt dem Rat
der Technischen Fakultät
der Albert-Ludwigs-Universität Freiburg

2015

von
M.Sc. Bioinformatiker (Univ.)
**Kousik Kundu**

**Dekan**
Prof. Dr. Georg Lausen
Databases and Information Systems
Department of Computer Science
University of Freiburg

**Vorsitz**
Prof. Dr. Christoph Scholl
Operating Systems
Department of Computer Science
University of Freiburg

**Beisitz**
Prof. Dr. Christian Schindelhauer
Computer Networks and Telematics
Department of Computer Science
University of Freiburg

**Datum der Promotion**
April 21, 2015

**Gutachter**
Prof. Dr. Rolf Backofen
Bioinformatics
Department of Computer Science
University of Freiburg

**Gutachter**
PD Dr. Björn Voß
Genetics & Exp. Bioinformatics
Institute for Biology III
University of Freiburg

***"Janani Janma-bhoomi-scha Swargadapi Gariyasi"***

(Devanagari: "जननी जन्मभूमिश्च स्वर्गादपि गरीयसी")

*"Mother and motherland are superior to Heaven"*

***Dedicated to my beloved parents …..***

# Acknowledgements

## *My sincere thanks to ......*

○ ○ ○ ○ ○ *First and foremost, I would like to express my deepest gratitude to Prof. Dr. Rolf Backofen for giving me the opportunity to pursue my PhD under his supervision. It was an absolute pleasure to work under his guidance as he allowed me to engage in exciting scientific discussions, offered me an excellent working platform, and provided me an extreme research freedom.*

○ ○ ○ ○ ○ *I would like to express my gratitude to PD Dr. Björn Voß for his interest in my thesis and kindness to review it. I would like to thank Prof. Dr. Christoph Scholl and Prof. Dr. Christian Schindelhauer for being a part of my PhD examination committee.*

○ ○ ○ ○ ○ *I would like to extend my gratitude to Dr. Farbizio Costa for helping me in every aspects of my research work. His expert guidance to understand the machine learning algorithms was invaluable. I am thankful to Prof. Dr. Michael Huber for his contribution mainly in the biological interpretation of SH2-peptide interactions.*

○ ○ ○ ○ ○ *I would like to thank my current and former lab members for their various help and providing an excellent working environment. The time I had spent with you during countless coffee/cake brakes, dart breaks, and social events will always be remembered. My heartiest thanks to Robert Kleinkauf, Christina Otto, Martin Mann, Dominic Rose, Omer Alkhnbashi, Andreas Richter, and Reelin Sinha for their generous help. Their support helped me to adjust in a completely new country, which is thousands of miles away from my home. They made me feel very comfortable at the lab and ensured me not to feel like a lonely outsider. And of course, I cannot forget to extend my thanks to Monika Degen-Hellmuth for her support in administrative stuffs.*

○ ○ ○ ○ ○ *I am very thankful to Deepti, Naveen, and Fabrizio for proofreading this thesis. My special thanks to Christina for her contribution in writing the "Zusammenfassung" part of this thesis.*

○ ○ ○ ○ ○ *I have been very lucky to have good friends in every stages of my life. They are like oxygen to me. I would like to take this opportunity to thank them all. My special thanks to those with whom I had spent my days in Freiburg. You made my life outside academics so easy and enjoyable. I had lot of fun with you all.*

○ ○ ○ ○ ○ *Last but not least, I would like to thank my beloved parents: Mahua Kundu and Asit Baran Kundu, my sweet brother: Anirban Kundu, my lovely wife: Deepti Jaiswal, and other family members. Without their unconditional love, endless support, and non-stop encouragement none of this would have been possible.*

# Contents

# Abstract

Protein-protein interactions (PPIs) are one of the most essential cellular processes in eukaryotes that control many important biological activities, such as signal transduction, differentiation, growth, cell polarity, apoptosis etc. Many PPIs in cellular signaling are mediated by modular protein domains. Peptide recognition modules (PRMs) are an important subclass of modular protein domains that specifically recognize short linear peptides to facilitate their biological functions. Hence, it is important to understand the intriguing mechanisms by which hundreds of modular domains specifically bind to their target peptides in a complex cellular environment. In recent years, an unprecedented progress has been made in high-throughput technologies to describe the binding specificities of a number of modular protein domain families. Therefore, given the high binding specificity of PRMs, *in silico* prediction of their cognate partners is of great interest.

In the first part of this thesis, we describe the main high-throughput technologies (microarray, phage display etc.) that are widely used for defining the binding specificity of PRMs. Currently, several computational methods have been published for the prediction of domain-peptide interactions. Here, we provide a comprehensive review on these methods and their applications. We also describe the major drawbacks (e.g., linearity problem, peptide alignment problem, data-imbalance problem etc.) of these existing tools that are successfully addressed in our study.

In the second part of this thesis, we present three methods for predicting domain-peptide interactions mediated by three diverse PRM families (i.e., SH2, SH3, and PDZ domain). In order to circumvent the linearity problem, our methods use efficient kernel functions, which exploit higher-order dependencies between amino acid positions. For the prediction of SH2-peptide interactions, polynomial kernels are used to train the classifiers. In addition, we show how to handle the data-imbalance problem by using an efficient semi-supervised technique. For the prediction of SH3-peptide interactions, graph kernels are used for training the classifiers. Graph kernel feature representation allows us to include the physico-chemical properties of each amino acid in the peptides, which increases the generalization capacity of the classifier. By using this kernel function, we were able to eliminate the need of an initial peptide alignment, since the alignment of proline-rich peptides targeted by SH3 domains is a hard task and an error-prone alignment can severely affect the predictive performance of the classifier. Moreover, we developed a generative approach for refining the confidence negative data. In the case of PDZ-peptide interactions, we cluster hundreds of PDZ domains from different organisms, i.e., human, mouse, fly, and worm, based on their binding specificity, and build a single comprehensive model for a set of multiple PDZ domains. In this way, we show that the domain coverage can be increased by using an accurate clustering technique. For training the classifier, a Gaussian kernel function is used. Similar to SH2-peptide interactions, a semi-supervised technique was applied to generate high-confidence negative data.

In the third part of this thesis, we describe the applications and performance evaluations of our methods. We compared our methods with several other existing tools and achieved a much higher performance, which was measured by sensitivity, specificity, precision, AUC PR, and AUC ROC. Our methods were further evaluated on various experimentally verified datasets and as a predictive result, they outperformed the state-of-the-art approaches. To uncover the novel and biologically relevant interactions, we performed a genome-wide prediction. Furthermore, a term-centric enrichment analysis has been performed to unveil the novel functionalities of the predicted interactions.

In the last part of this thesis, we introduce a new and efficient web server, which contains three tools (i.e., `SH2PepInt`, `SH3PepInt`, and `PDZPepInt`), for the prediction of modular domain-peptide interactions. Currently, we offer 51 and 69 single domain models for SH2 and SH3 domains, respectively, and 43 multiple domain models, which cover 227 domains, for PDZ domains across several organisms.

In summary, this thesis presents machine learning methods for predicting the binding peptides of three diverse PRM families where the training data was derived from various high-throughput experiments. Most importantly, this thesis addresses the major computational challenges in the field of modular domain-peptide interactions. We offer the largest set of models to date for the prediction of modular domain mediated interactions.

# Zusammenfassung

Protein-Protein-Interaktionen (PPIs) zählen mit zu den wesentlichen Prozessen in Eukaryoten, die viele wichtige biologische Vorgänge (wie Signaltransduktion, Differenzierung, Wachstum, Zellpolarität, Apoptose usw.) kontrollieren. Viele PPIs in der Zellkommunikation werden durch modulare Proteindomänen vermittelt. Peptiderkennungsmodule (PRMs) sind eine wichtige Unterklasse der modularen Proteindomänen, die spezifisch kurze lineare Peptide erkennen, um ihre biologischen Funktionen zu ermöglichen. Demzufolge ist es wichtig, die faszinierenden Mechanismen zu verstehen, durch die hunderte modulare Domänen in einer komplexen zellulären Umgebung spezifisch an ihre Zielpeptide binden. In den vergangenen Jahren wurde ein noch nie da gewesener Fortschritt in Hochdurchsatz-Technologien gemacht, um die Bindungsspezifität einer Reihe von Familien modularer Proteindomänen zu beschreiben. Aus diesem Grund sind wegen der hohen Bindungsspezifität von PRMs *in silico* Vorhersagen ihrer spezifischen Partner von großem Interesse.

Im ersten Teil dieser Arbeit beschreiben wir die wichtigsten Hochdurchsatz-Technologien (Microarray, Phagen-Display usw.), die weithin verwendet werden, um die Bindungsspezifität von PRMs zu bestimmen. Gegenwärtig wurden etliche computergestützte Methoden für die Vorhersage von Domäne-Peptid-Interaktionen veröffentlicht. Wir stellen einen umfassenden Überblick über diese Methoden und ihre Anwendungen bereit. Wir beschreiben auch die bedeutendsten Nachteile (zum Beispiel Linearitäts-Problem, Peptid-Alignment-Problem, Daten-Ungleichgewichts-Problem usw.) dieser bestehenden Methoden, die wir in unserer Studie erfolgreich angehen werden.

Im zweiten Teil dieser Arbeit stellen wir drei Methoden zur Vorhersage von Domäne-Peptid-Interaktionen dar, die durch drei unterschiedliche PRM-Familien (d.h. SH2-, SH3- und PDZ-Domänen) vermittelt werden. Um das Linearitäts-Problem zu umgehen, verwenden unsere Methoden effiziente Kernel-Funktionen, die Abhängigkeiten höherer Ordnung zwischen Positionen einer Aminosäure ausnutzen. Für die Vorhersage von SH2-Peptid-Interaktionen werden polynomielle Kernel verwendet, um die Klassifikatoren zu trainieren. Zusätzlich zeigen wir, wie das Daten-Ungleichgewichts-Problem gehandhabt werden kann, indem ein effizientes semi-überwachtes Verfahren angewendet wird. Bei der Vorhersage von SH3-Peptid-Interaktionen werden Graph-Kernel für das Training der Klassifikatoren verwendet. Die Merkmalsrepräsentation des Graph-Kernels erlaubt es uns physikalisch-

chemische Eigenschaften jeder Aminosäure in den Peptiden einzubeziehen, was das Verallgemeinerungsvermögen des Klassifikators erhöht. Durch die Verwendung dieser Kernel-Funktion konnten wir die Erfordernis eines initialen Peptid-Alignments streichen. Dies ist besonders wichtig, da das Alignment von Prolin-reichen Peptiden, die ein Ziel von SH3-Domänen sind, eine schwierige Aufgabe darstellt und ein fehlerhaftes Alignment die Vorhersage-Güte des Klassifikators schwerwiegend in Mitleidenschaft ziehen kann. Darüber hinaus entwickelten wir einen generativen Ansatz, um die sicher negativen Daten zu verfeinern. Im Falle der PDZ-Peptid-Interaktionen gruppieren wir hunderte PDZ-Domänen unterschiedlicher Organismen, d.h. Mensch, Maus, Fliege und Wurm, basierend auf ihrer Bindungsspezifität und erstellen ein einziges umfassendes Modell für eine Reihe multipler PDZ-Domänen. Auf diese Weise zeigen wir, dass die Abdeckung der Domänen durch die Verwendung eines exakten Clusterbildungsansatzes erhöht werden kann. Um den Klassifikator zu trainieren, wird eine Gauß-Kernel-Funktion verwendet. Ähnlich wie bei SH2-Peptid-Interaktionen wurde ein semi-überwachter Ansatz eingesetzt, um die sicher negativen Daten zu generieren.

Im dritten Teil der Arbeit beschreiben wir die Anwendungen und die Leistungsbewertung unserer Methoden. Wir haben unsere Methoden mit mehreren anderen existierenden Programmen verglichen und erreichten eine wesentlich höhere Leistungsfähigkeit, die durch Sensitivität, Spezifität, Genauigkeit, AUC PR und AUC ROC gemessen wurde. Unsere Methoden wurden darüber hinaus auf verschiedenen experimentell abgesicherten Datensätzen ausgewertet und als Vorhersage konnten sie dem Stand der Technik entsprechende Ansätze übertreffen. Um die neuartigen und biologisch relevanten Interaktionen aufzudecken, führten wir eine genomweite Vorhersage durch. Zusätzlich wurde eine “term-centric enrichment analysis” durchgeführt, um die neuartigen Funktionsweisen der vorhergesagten Interaktionen zu enthüllen.

Im letzten Teil dieser Arbeit präsentieren wir einen neuen und effizienten Web-Server, der drei Tools (d.h. `SH2PepInt`, `SH3PepInt` und `PDZPepInt`) für die Vorhersage von modularen Domäne-Peptid-Interaktionen beinhaltet. Derzeit bieten wir 51 bzw. 69 einzelne Domänen-Modelle für SH2- und SH3-Domänen an, und 43 multiple Domänen-Modelle, die 227 Domänen umfassen, für PDZ-Domänen mehrerer Organismen.

Zusammenfassend stellt diese Arbeit maschinelle Lernverfahren für die Vorhersage der gebundenen Peptide von drei unterschiedlichen PRM-Familien dar, wobei die Trainingsdaten von zahlreichen Hochdurchsatz-Experimenten stammten. Am Bedeutsamsten ist, dass sich diese Arbeit mit den großen rechnergestützten Herausforderungen im Bereich der modularen Domäne-Peptid-Interaktionen befasst. Wir bieten die bislang größte Menge an Modellen für die Vorhersage von Interaktionen, die durch modulare Domänen vermittelt werden.

*Specificity tree of human PDZ domains*

# List of publications

**This thesis is based on the following publications:**

[P1] **Kousik Kundu***, Martin Mann*, Fabrizio Costa, and Rolf Backofen. MoDPepInt: an interactive web server for prediction of modular domain-peptide interactions. *Bioinformatics*, 2014.

[P2] **Kousik Kundu** and Rolf Backofen. Cluster based prediction of PDZ-peptide interactions. *BMC Genomics*, 15 Suppl 1 pp. S5, 2014.

[P3] **Kousik Kundu***, Fabrizio Costa*, and Rolf Backofen. A graph kernel approach for alignment-free domain-peptide interaction prediction with an application to human SH3 domains. *Bioinformatics*, 29 no. 13 pp. i335-i343, 2013.

[P4] **Kousik Kundu**, Fabrizio Costa, Michael Huber, Michael Reth, and Rolf Backofen. Semi-Supervised Prediction of SH2-Peptide Interactions from Imbalanced High-Throughput Data. *PLoS One*, 8 no. 5 pp. e62732, 2013.

* Joint first authors

# Chapter 1

# Introduction

## 1.1. Motivation

*"Arise! Awake! and stop not until the goal is reached"* –Swami Vivekananda

It has been more than 30 years since the concept of signal transduction emerged. The discovery of various signal transduction pathways opened a new door for an in-depth understanding of cell signaling. Signal transduction is basically the lens through which we examine all cellular activities. This process is initiated with the acceptance of extracellular signals by the receptor proteins at plasma membrane, then these signals get transduced inside the cell interior and finally they take control of numerous multi-protein complexes to regulate variety of signaling pathways [1]. In the last decade, the fundamental knowledge of signal transduction has been successfully translated to the clinic. There have been some remarkable achievements in terms of therapeutic targets, particularly in cancer [2]. Two drugs, i.e., imatinib and trastuzumab, are widely used for the treatment of chronic myelogenous leukemia (CML) and breast cancers, respectively. Other signal transduction inhibitors are currently in advanced clinical trials.

The 1980s were an exciting time for cell signaling research. The path breaking discovery of modular protein domains and their role in the regulation of signaling pathways was an unprecedented event in biological science that drastically changed our conceptual understanding of protein function. Tony Pawson and co-workers first discovered a non-catalytic conserved modular domain, namely Src homology 2 (SH2), which regulates various signal transduction pathways, and set the stage for exciting discoveries of other modular domains in subsequent years [3]. Importantly, modular domains play an important role in the therapeutic development. For example, SH2 domains can serve as biomarkers for normal or perturbed signaling networks and can be used for *personalized medicine* [4]. These domains specifically bind to their cognate partners to facilitate their molecular functions. However, the detailed mechanism by which thousands of modular protein domains target their binding partners with high specificity is an open challenge with high relevance.

The data generated by various high-throughout techniques seems to be a perfect source for investigating the molecular functions of the modular protein domains. In recent years, a monumental progress has been made in the field of high-throughput technologies. Large amount of data is being generated by various high-throughput techniques to address the binding specificity of modular domains. Now the question is, how to exploit the wealth of biological information hidden in these huge amount of data. Here, computer science plays a crucial role to handle these data and to make sense out of it. In modern research, application of computer science is indispensable to solve several complex biological problems.

Currently, several computational approaches, which use high-throughput data, have been published for the prediction of modular domain mediated interactions. However, these approaches have several shortcomings, starting from limited coverage, to restrictive modeling assumptions, to high computational complexity. The motivation of our work was to address these shortcomings. Therefore, we have resorted to a machine learning approach to accurately predict modular domain mediated interactions. Moreover, our intention was to build a tool that can be easily used by the biologists and thus we offer an easy-to-use web server, which will help biologists to pursue their research.

## 1.2. General overview

### 1.2.1. Outline of the thesis

In this thesis, three methods for predicting modular domain mediated interactions have been described. This thesis also addresses the open questions regarding computational prediction of domain-peptide interactions. The thesis is divided into six chapters as described below.

- **Chapter 1:** This introductory chapter describes the biological background of the work. Detailed description of the modular domains and their roles in cellular signaling will help the reader to understand the importance of the study.

- **Chapter 2:** This chapter gives a review on existing methods that includes high-throughput techniques and computational prediction methods, and their limitations for defining the specificity of the modular domains.

- **Chapter 3:** This chapter elucidates new prediction strategies for identifying the modular domain-peptide interactions for three different domains, i.e., SH2, SH3, and PDZ. These prediction methods are further divided into subsections.

- **Chapter 4:** Application and the performance evaluation of the proposed methods, and predicting novel interactions and their biological insights via genome-wide analysis have been described in this chapter.

- **Chapter 5:** This chapter presents `MoDPepInt`, a simple and easy-to-use web server for predicting modular domain-peptide interactions.

- **Chapter 6:** The final chapter concludes the proposed work and provides ideas that can open a new gateway for the future research in this field.

Each chapter starts with an overview, which will help readers to understand the gist of the chapter. A chapter specific (if not otherwise section specific) discussion is included for summarizing the particular part of the thesis. All supplementary materials, a list of abbreviations, and a plagiarism declaration have been mentioned in the appendix.

Here, I would like to state that all the work, which has been presented in this thesis, is my original research work and was published in four peer-reviewed journals [P1], [P2], [P3], and [P4] where I was the first author or one of the joint first authors. In section 1.2.2, contribution of other authors are stated. The aforementioned publications are recycled and/or reused in some parts of this thesis with appropriate references. Note that these publications were published under the Creative Commons Attribution license where authors retain the ownership of the copyright of their articles. This license also allows articles to be reused, modified, and distributed in any format, given that the original author and source are cited.

### 1.2.2. Statement of contributions

First, to mention my own contribution, I have done the major and most important parts of the work, which involve conceiving the work, data collection, designing and performing experiments, developing models, and writing manuscripts. In some cases, the work was entirely done by myself, e.g., [P2]. However, scientific collaboration is indispensable in the modern days of research. With no exception, this thesis also includes various collaborative works. Although I was the main contributor of the proposed work, scientists from internal and external groups also contributed by sharing ideas, expert biological knowledge, and fruitful discussions. In particular, Dr. Fabrizio Costa helped me to understand the machine learning algorithms, and guided me for [P3] and [P4]. Along with his contribution in writing some parts of the aforementioned publications, he has also significantly contributed in the method section of [P3]. Dr. Martin Mann mainly implemented the `MoDPepInt` web server [P1]. Prof. Dr. Michael Huber partly contributed to the biological part of the [P4]. Christina Otto has contributed in German translation of the “abstract (Zusammenfassung)” of this thesis. And all of my work was supervised by Prof. Dr. Rolf Backofen. Therefore, I have chosen *“we”* as an appropriate pronoun instead of *“I”* and have used throughout the thesis.

## 1.3. Cell signaling and signal transduction

Cell signaling is a complex biological mechanism through which a cell receives extracellular signals, processes those signals, and responds accordingly. By this mechanism, one cell communicates with others, and controls numerous cellular activities. The extracellular signal molecules produced by various cells are the most essential for cell communication. However, these signal molecules are not sufficient to transfer the information to a specific cell; they need a set of receptor proteins from each cell that receive the incoming information and transfer it into the intracellular environment. There are three main families of cell surface receptors: (i) ion-channel-linked receptors, (ii) G-protein-linked receptors, and (iii) enzyme-linked receptors, and they all react to the extracellular signals in a different way [1]. Once these receptor proteins get activated by the extracellular signal molecules, such as hormones, neurotransmitters etc., they subsequently bind to a series of intracellular signaling proteins to relay the signals into cell interior (Figure 1.1). The intracellular signaling proteins have a variety of functions, including distribution and amplification of the signals. Some intracellular proteins also integrate signals from other signaling pathways [1]. This whole process is known as signal transduction where the extracellular signals are transduced into a cellular environment by the cell surface receptor proteins; henceforth, activate a cascade of signaling pathways inside the cell, and eventually, trigger the functional changes of the cell.

Phosphorylation is an important post-translational modification of a protein that plays a crucial role in the regulation of signal transduction pathways. In this process, an amino acid is phosphorylated by a protein kinase. Phosphorylation normally occurs on a serine, a threonine, or a tyrosine residue, although phosphorylation of basic amino acids has also been observed [5]. Signaling proteins often serve as *"molecular switches"*, which are activated by protein phosphorylation. Receptor tyrosine kinases (RTKs) are the largest kinase family that phosphorylate specific tyrosine residues in a protein and play a vital role in signal transduction by regulating a variety of essential cellular processes, such as proliferation, differentiation, growth, migration, apoptosis, and malignant transformation in metazoans [6–9]. Many of the signaling proteins contain modular protein domains that mainly control the function of complex protein assemblies and regulate signal transduction pathways.

## 1.4. Modular protein domains and their binding specificity

Protein-protein interaction (PPI) is a major area of biological science to understand the transduction of cellular signals. PPIs take an indispensable role to transfer information in signal transduction pathways. The regulation of numerous signal transduction pathways are mainly mediated by the binding of a modular protein domain with a short linear peptide [10]. These peptide recognition modular protein domains are also known as peptide recognition modules (PRMs). Figure 1.1 is a hypothetical illustration of signal transduction pathways

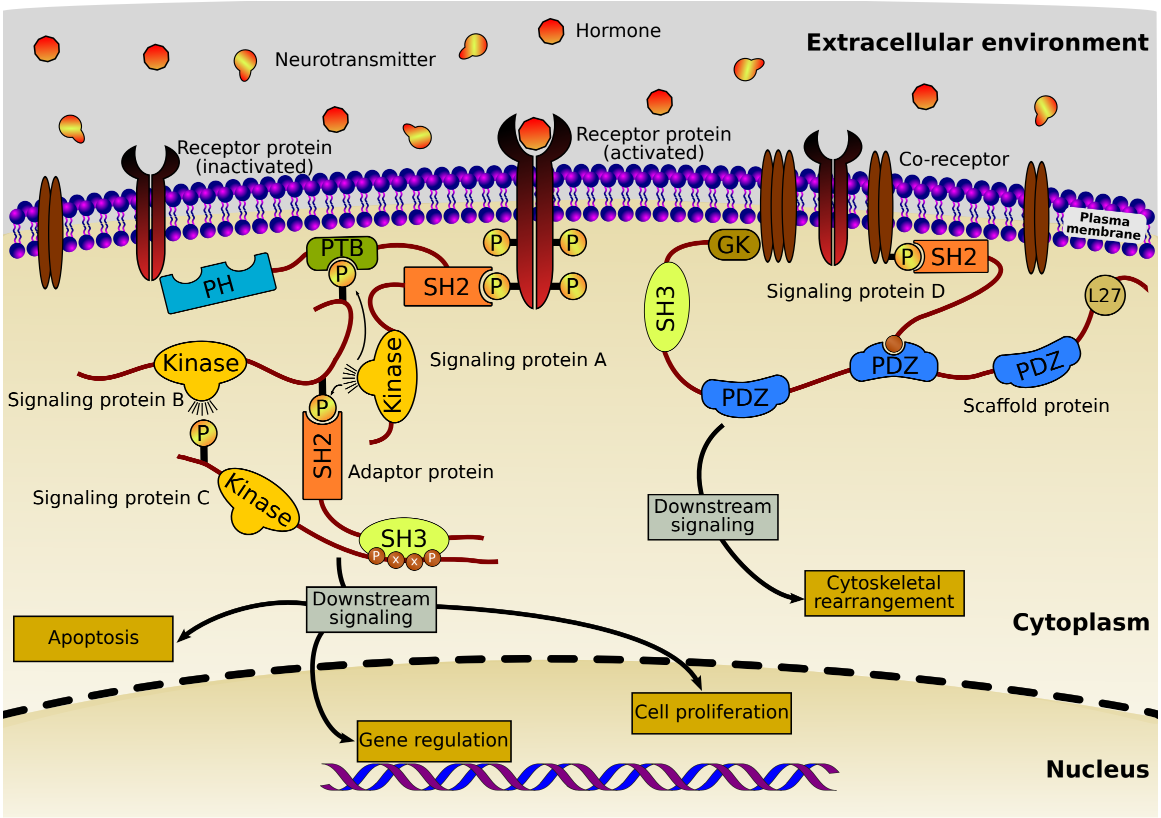


**Figure 1.1.:** A hypothetical illustration of signal transduction processes. In this figure, when a receptor protein interacts with a growth factor, it gets phosphorylated and then binds to the SH2 domain of signaling protein A. Then the kinase domain of protein A phosphorylates two tyrosine residues of signaling protein B, which bind to PTB domain of protein A and SH2 domain of an adaptor protein, respectively. Signaling protein C is then phosphorylated by the kinase domain of protein B, and its proline-rich region is targeted by the SH3 domain of the adaptor protein. On the other hand, a phosphorylated co-receptor is targeted by SH2 domain of signaling protein D. A PDZ domain of a scaffold protein then targets the C-terminal tail of protein D. In both cases, the further downstream signaling processes proceed with the analogous fashion and eventually, as a result, alter the function of the cell.

that are mediated by modular protein domains. There are hundreds of PRMs spread into human proteome, and they have highly specific binding preferences. For example, Src homology 2 (SH2) and phosphotyrosine binding (PTB) domains recognize peptides containing a phosphorylated tyrosine (pTyr) residue, 14-3-3 domains bind to phosphoserine (pSer) containing peptides, Src homology 3 (SH3) and WW domains recognize protein-rich peptide motifs, Eps15 homology (EH) domains bind to the NPF motif containing peptides, and PSD-95/DLG1/ZO-1 (PDZ) domains recognize the C-terminal peptide tails of the binding proteins (see Figure 1.2) [11]. In this thesis, we focused on three different modular domains, SH2, SH3, and PDZ, that are essential in various cellular processes.

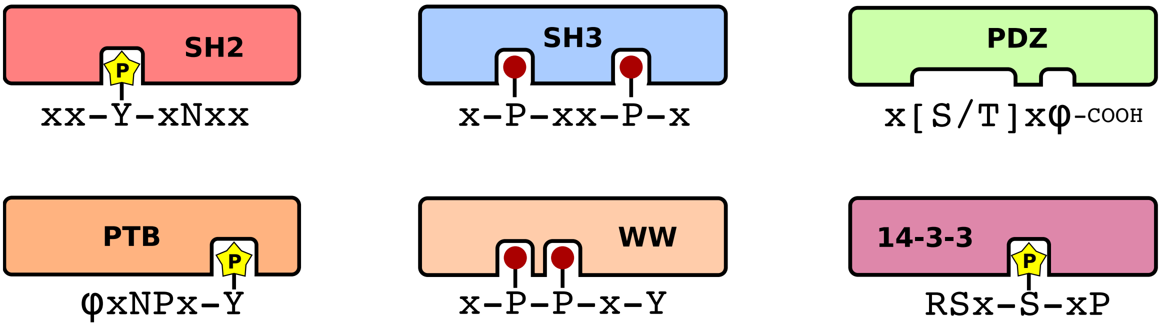


**Figure 1.2.:** Canonical binding specificity of various well-studied peptide recognition modules. In the motif, x represents natural amino acids and $\phi$ represents hydrophobic amino acids. The "P" symbol with yellow background represents the phosphorylation.

### 1.4.1. SH2 domains

SH2 domains are the largest family of peptide recognition modules (PRMs), normally found in intracellular signal transducing proteins [12–14]. In 1986, Tony Pawson and his colleagues first identified the SH2 domain from oncogenic v-FPS/FES cytoplasmic tyrosine kinase in Fujinami sarcoma virus [3]. In subsequent years, other SH2 domains, such as v-CRK, PLC$\gamma$1, and RasGAP, were discovered [15, 16]. Since then hundreds of SH2 domains have been found across the eukaryotic species, however, they are more abundant in metazoans [17, 18]. Currently, 122 SH2 domains from 112 unique human proteins have been reported in the `UniProtKB/Swiss-Prot` database, release 2015-01 [19]. SH2 domains are identified in wide range of signaling proteins, including protein kinases, protein phosphatases, adaptor proteins, scaffold proteins, transcription factors, signal regulator proteins [6]. Based on the composition of modular domains, SH2 containing proteins were classified into 11 functional categories by Liu *et al.* [20]. Figure 1.3 illustrates the phylogenetic tree of all SH2 domains with their functional annotation. SH2 domains are known to specifically recognize phosphorylated tyrosine (pTyr) residues and mediate intracellular signaling [21, 22]. Researches using the peptide libraries have shown that each SH2 domain binds with a specific subset of phosphopeptides [23–26]. Cytoplasmic protein tyrosine kinases (PTKs) and receptor tyrosine kinases (RTKs) play a central role to facilitate numerous cellular processes by interacting with SH2 domains [6]. For example, a well-studied receptor tyrosine kinase, namely epidermal growth factor receptor (EGFR), mediates intercellular communication to regulate wide range of cellular activities by interacting with SH2 domains [6, 7]. There are some evidences that mutations in some SH2 domains can cause several human diseases, such as XLP syndrome [27], Noonan syndrome [28], X-linked $\alpha$-gammaglobulinemia [29], and basal cell carcinoma [30].

#### Sequence and structure

SH2 domains are approximately 100 amino acids in length, and are structurally conserved protein domains that comprise a central anti-parallel $\beta$ sheets flanked by two $\alpha$ helices, and

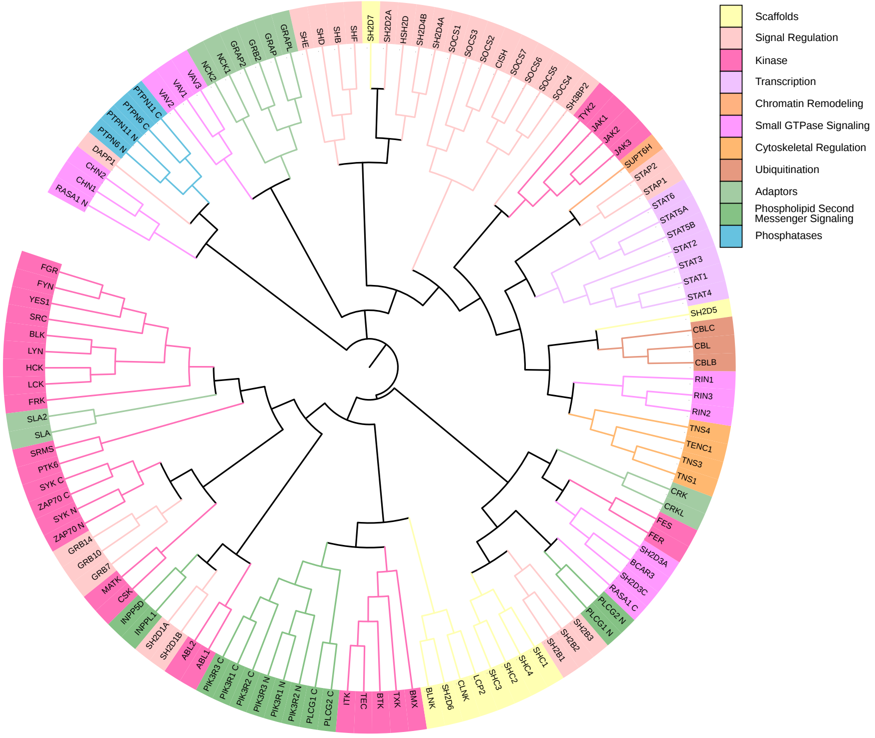
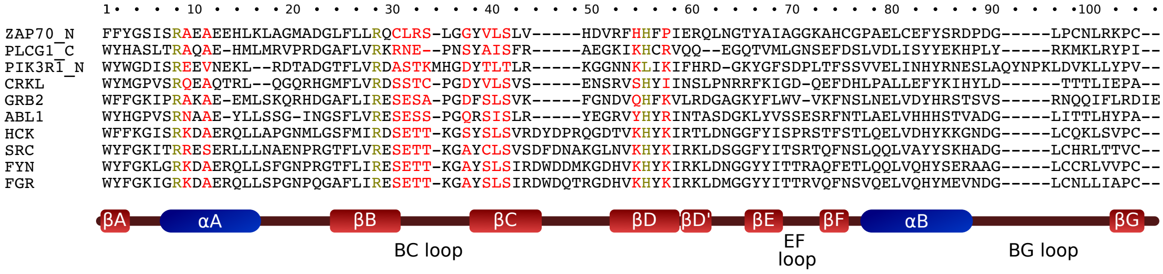


**Figure 1.3.:** Phylogenetic tree of all 122 human SH2 domains available in `UniProtKB/Swiss-Prot` database, release 2015-01 [19]. A total number of 11 functional classes derived from [20] are presented in different colors. The tree was built by `ClustalW` [31], and `iTOL` software was used for the visualization [32].

an additional tripled-stranded $\beta$ sheets on the C-terminus (see Figure 1.5.A) [33]. Almost 20 years ago, the first SH2-peptide complex structure was solved by Waksman *et al.*, which unveiled a high affinity interaction between SRC SH2 domain and a phosphorylated peptide motif, PQ-pY-EEIP [34]. Recent study revealed that there are approximately 70 experimentally validated structures of unique SH2 domains available in the `PDB` database [35]. SH2 domains contain an evolutionary conserved phosphopeptide binding pocket formed by some conserved residues, which interacts with their target peptides [12]. Another binding pocket, namely specificity pocket, has also been observed, which is mainly formed by $\beta$D, $\beta$E, and the loop regions of SH2 domains. The sequence and structural organization of SH2 domains are illustrated in Figure 1.4. Although residues in this specificity pocket are less conserved than the rest of the domain, it forms a conserved structure that binds to C-terminal residues of the binding peptides; this structural conservation has been seen from early invertebrate to human [18].

### Binding specificity

Previous studies showed that each SH2 domain distinctly interacts with phosphotyrosine (pTyr) containing peptides [18, 20]. The negatively charged phosphate moiety on the tyrosine residue specifically targeted by the phospho-binding pocket of the SH2 domains. The



```
         1• • • •10 • • • •20 • • • •30 • • • •40 • • • •50 • • • •60 • • • •70 • • • •80 • • • •90 • • • •100 • • •
ZAP70_N  FFYGSISRAEAEEHLKLAGMADGLFLLRQCLRS-LGGYVLSLV-----HDVRFHHFPIERQLNGTYAIAGGKAHCGPAELCEFYSRDPDG-----LPCNLRKPC--
PLCG1_C  WYHASLTRAQAE-HMLMRVPRDGAFLVRKRNE--PNSYAISFR-----AEGKIKHCRVQQ--EGQTVMLGNSEFDSLVDLISYYEKHPLY-----RKMKLRYPI--
PIK3R1_N WYWGDISREEVNEKL--RDTADGTFLVRDASTKMHGDYTLTLR-----KGGNNKLIKIFHRD-GKYGFSDPLTFSSVVELINHYRNESLAQYNPKLDVKLLYPV--
CRKL     WYMGPVSRQEAQTRL--QGQRHGMFLVRDSSTC-PGDYVLSVS-----ENSRVSHYIINSLPNRRFKIGD-QEFDHLPALLEFYKIHYLD------TTTLIEPA--
GRB2     WFFGKIPRAKAE-EMLSKQRHDGAFLIRESESA-PGDFSLSVK-----FGNDVQHFKVLRDGAGKYFLWV-VKFNSLNELVDYHRSTSVS-----RNQQIFLRDIE
ABL1     WYHGPVSRNAAE-YLLSSG-INGSFLVRESESS-PGQRSISLR-----YEGRVYHYRINTASDGKLYVSSESRFNTLAELVHHHSTVADG-----LITTLHYPA--
HCK      WFFKGISRKDAERQLLAPGNMLGSFMIRDSETT-KGSYSLSVRDYDPRQGDTVKHYKIRTLDNGGFYISPRSTFSTLQELVDHYKKGNDG-----LCQKLSVPC--
SRC      WYFGKITRRESERLLLNAENPRGTFLVRESETT-KGAYCLSVSDFDNAKGLNVKHYKIRKLDSGGFYITSRTQFNSLQQLVAYYSKHADG-----LCHRLTTVC--
FYN      WYFGKLGRKDAERQLLSFGNPRGTFLIRESETT-KGAYSLSIRDWDDMKGDHVKHYKIRKLDNGGYYITTRAQFETLQQLVQHYSERAAG-----LCCRLVVPC--
FGR      WYFGKIGRKDAERQLLSPGNPQGAFLIRESETT-KGAYSLSIRDWDQTRGDHVKHYKIRKLDMGGYYITTRVQFNSVQELVQHYMEVNDG-----LCNLLIAPC--
```

**Figure 1.4.:** Alignment and structural organization of human SH2 domains. Colored residues (red and green) are the main specificity determinants for the SH2 domains. They are responsible for forming the phosphopeptide binding pocket [41]. Residues in green specifically bind to the negatively charged phosphorylated tyrosine (pTyr) residue of the binding peptide [36]. The `MUSCLE` program was used for the alignment [42].

surface residues of the SH2 domains, mainly a highly conserved Arg at position 5 of $\beta$B, an Arg at position 2 of $\alpha$A, and a His at position 4 of $\beta$D play the major role to form the binding pocket [36]. Figure 1.4 illustrates the surface residues that are responsible for making contacts with pTyr containing peptides. However, the specificity of a SH2 domain can be altered by engineering the surface loops [37]. The Arg residue at position 5 of $\beta$B is occurs in most of the SH2 domains (118 out of 121), except RIN2 and TYK, which contain a His, and SH2D5, which contains a Trp [20]. It has been shown that the mutation of Arg ($\beta$B5) or His ($\beta$D4) can abolish the pTyr dependent interactions [38].

A global position system has been induced where the phosphotyrosine (pTyr) residue is given position 0, N-terminal residues with respect to the pTyr are given position $-1$, $-2$, and so on, and C-terminal residues with respect to the pTyr are given position $+1$, $+2$, and so on. Although the pTyr residue is essential in most of the SH2-peptide interactions, the binding specificity of a given SH2 domain is determined by the C-terminal residues to the pTyr, particularly from $+1$ to $+5$ [21, 36]. For example, CRK SH2 domains have a very strong preference for a Leu or Pro residue at position $+3$ (xx-pY-xx[L/P]x, where x represents any naturally occurring amino acid) in the binding motifs. Similarly, GRB2 strongly prefers an Asn at position $+2$ (xx-pY-xNxx) and BRDG1 prefers a Leu at position $+4$ (xx-pY-xxxL) [39]. Although C-terminal residues with respect to pTyr are the most important to define the binding specificity of SH2 domains, N-terminal residues have also been identified to be targeted by some SH2 domains. For example, a hydrophobic residue at position $-2$ ($\Phi$x-pY-xxxx, where $\Phi$ represents a hydrophobic residue) is preferred by the SH2 domain from PTPN11 protein [40].

Since SH2 domains have distinct binding preferences, the contextual sequence information of the binding peptides is the key to discriminate among the binding specificity of different SH2 domains [43]. Although SH2 domains from the ABL, CRK, BRK, VAV, and RASA1 proteins have a basic preference to interact with a peptide, xx-pY-xx[P/L]x, they bind to a subset of peptides that contains this motif [36]. The underneath mechanism of this intriguing nature is due to the presence and absence of permissive and non-permissive amino

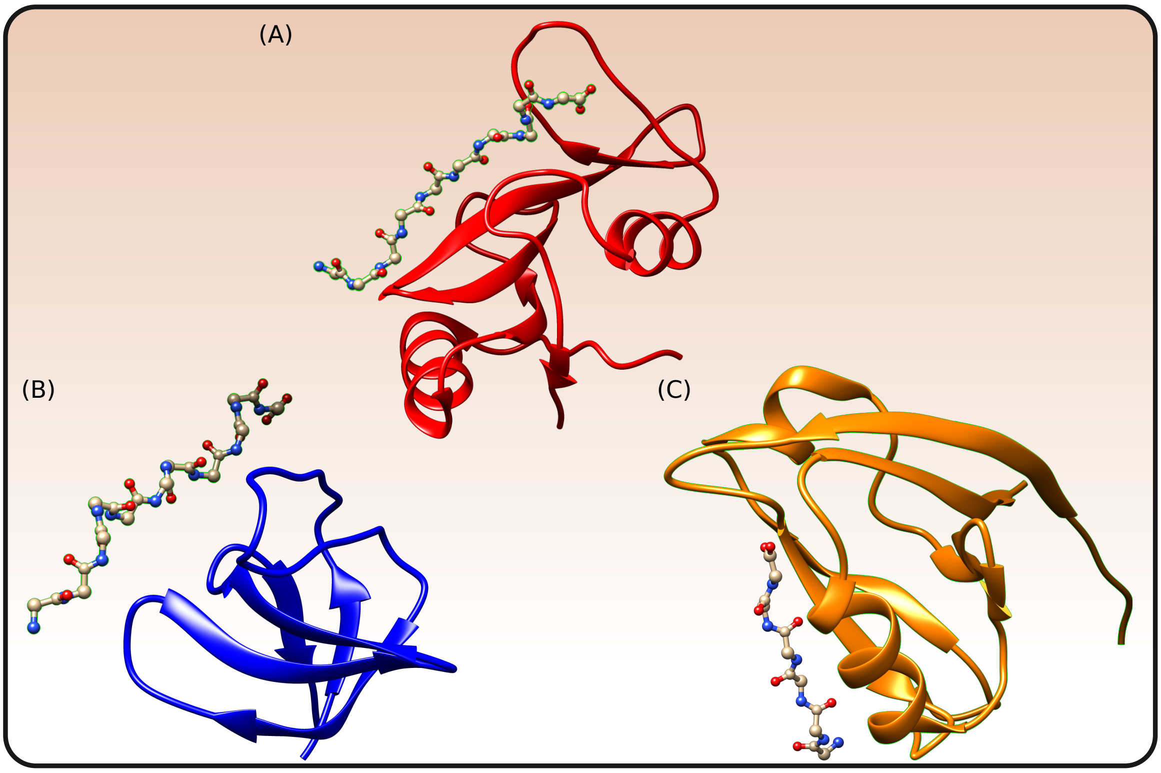


**Figure 1.5.:** Domain-peptide complex 3D structures for SH2, SH3, and PDZ domains. (A) SH2-peptide complex structure (PDB id: 1D4W), (B) SH3-peptide complex structure (PDB id: 2BZ8), and (C) PDZ-peptide complex structure (PDB id: 4G69). UCSF Chimera was used for the visualization [48].

acids in the binding peptides where permissive residues promote the interaction and the non-permissive residues inhibit or diminish the interaction. For example, while an Arg at +4 to the pTyr is a permissive factor for the CRK interaction, it is non-permissive and thus rejects the BRK interaction. Similarly, while a Glu at +1 to the pTyr promotes the BRK interaction, it prohibits the CRK interaction (see Figure 2.3 for details) [36, 43]. Therefore, accurately identifying the specific binding peptides for a given SH2 domain is indispensable to determine its biological function.

Interestingly, SH2 domains are also found to interact with non-phosphorylated peptide motifs, however, the binding affinity is 1000-fold less than the SH2-pTyr peptide interactions [33, 34, 44]. One of the well-studied examples for a phospho-independent interaction is the interaction between the SH2D1A/SAP SH2 domain and the SLAM receptor protein, albeit the binding affinity was notably low [26]. In this case, the SH2 domain from SH2D1A protein does not need a pTyr residue for the interaction. Other examples where SH2 domains, such as TENC1, SHC, and CTEN, bind to their cognate partners in a phospho-independent manner have also been reported previously [45–47].

### 1.4.2. SH3 domains

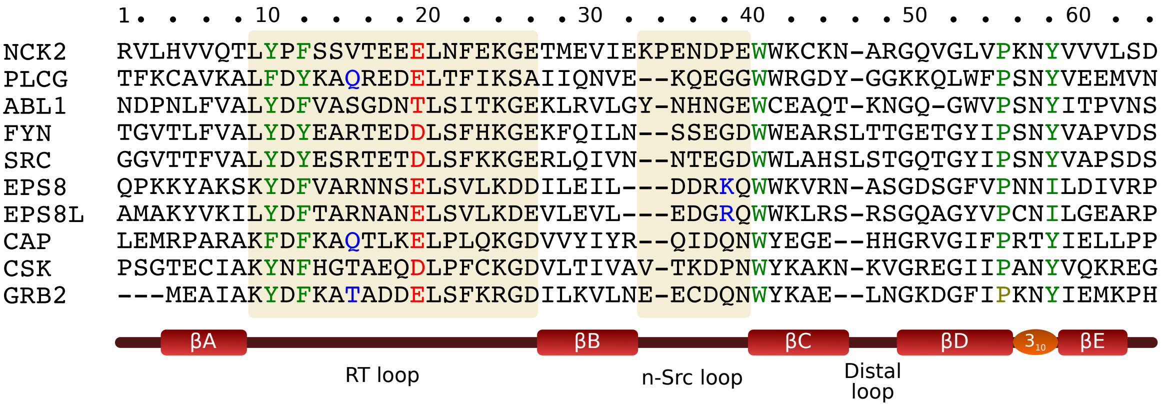
SH3 domains are characterized as an important class of peptide recognition module (PRM) that specifically recognize short linear proline-rich peptide sequences and play a pivotal role in a wide range of cellular processes, such as intracellular signaling, growth, cytoskeletal rearrangements, cell-communication, cell movement, differentiation etc. [49–51]. In 1988, SH3 domains were first identified in eukaryotes [52, 53]. It is known that SH3 domains are most abundant in eukaryotic genomes, but interestingly, the occurrence of these domains in various genomes corresponds with the genome complexity. For example, 28, 90, and 300 SH3 domains are encoded in yeast, drosophila, and human genomes, respectively [54]. Previous study showed that the binding affinity of SH3 domains can be increased up to 40-fold by directed evaluation [55]. Along with its occurrence in multiple copies in a protein, it can also occur with other modular domains, such as SH2, PDZ etc. For example, two SH3 domains and an SH2 domain are found in GRB2 protein while an SH3 domain along with three PDZ domains are encoded in DLG1 protein. Moreover, some SH3 domains were found to bind their physiological partners through tertiary contacts instead of targeting a defined sequence motif [56].

#### Sequence and structure

SH3 domains are one of the small modular domains found in signaling proteins that are mainly involved in signal transduction, membrane trafficking, cytoskeleton organization etc. [49]. SH3 domains are typically 60-70 amino acids in length. Although the sequence similarity of two SH3 domains is only 25%, these domains are structurally very conserved [57]. SH3 domains are composed of a conserved $\beta$-barrel fold, which is formed by $5-6$ $\beta$ strands arranged in two anti-parallel $\beta$ sheets (see Figure 1.5.B). These $\beta$ strands are connected by some structural conformations, such as an RT-loop, an n-Src loop, a $3_{10}$ helix, and a distal loop [58]. In most of the SH3 domains, the RT loop and the n-Src loop are generally 18 and 4 amino acids long, respectively. However, the length of these loops can be varied greatly for some SH3 domains. For example, the length of the RT loop and the n-Src loop can be ranging from $15-31$ and $3-31$ amino acids, respectively [59]. These loops are highly specific to recognize their cognate partners. It has been observed that variations of these loops can produce a new SH3 domain with a novel binding specificity [60, 61]. Sequence and structural organization of SH3 domains are depicted in Figure 1.6.

#### Binding specificity

SH3 domains are probably the most widespread protein domain found in protein databases. Since 25% of human proteins contain proline-rich regions [62] and SH3 domains recognize proline-rich peptides, it is an open challenge to understand how the hundreds of SH3 domains achieve a high specificity in selecting their physiological partners to regulate specific biological functions. The proline-rich peptide motifs recognized by most of the human SH3



```
       1 . . . . 10 . . . . 20 . . . . 30 . . . . 40 . . . . 50 . . . . 60 . .
NCK2   RVLHVVQTLYPFSSVTEEELNFEKGETMEVIEKPENDPEWWKCKN-ARGQVGLVPKNYVVVLSD
PLCG   TFKCAVKALFDYKAQREDELTFIKSAIIQNVE--KQEGGWWRGDY-GGKKQLWFPSNYVEEMVN
ABL1   NDPNLFVALYDFVASGDNTLSITKGEKLRVLGY-NHNGEWCEAQT-KNGQ-GWVPSNYITPVNS
FYN    TGVTLFVALYDYEARTEDDLSFHKGEKFQILN--SSEGDWWEARSLTTGETGYIPSNYVAPVDS
SRC    GGVTTFVALYDYESRTETDLSFKKGERLQIVN--NTEGDWWLAHSLSTGQTGYIPSNYVAPSDS
EPS8   QPKKYAKSKYDFVARNNSELSVLKDDILEIL---DDRKQWWKVRN-ASGDSGFVPNNILDIVRP
EPS8L  AMAKYVKILYDFTARNANELSVLKDEVLEVL---EDGRQWWKLRS-RSGQAGYVPCNILGEARP
CAP    LEMRPARAKFDFKAQTLKELPLQKGDVVYIYR--QIDQNWYEGE--HHGRVGIFPRTYIELLPP
CSK    PSGTECIAKYNFHGTAEQDLPFCKGDVLTIVAV-TKDPNWYKAKN-KVGREGIIPANYVQKREG
GRB2   ---MEAIAKYDFKATADDELSFKRGDILKVLNE-ECDQNWYKAE--LNGKDGFIPKNYIEMKPH
```

**Figure 1.6.:** Alignment and secondary structural elements of human SH3 domains. Residues occur in RT loop and n-src loop are represented in colored background. Variations in these loops largely determine the binding specificity of SH3 domains. Most conserved residues, that mainly participate in forming the PPII binding pockets, are shown in green. Residues, which are main specificity determinants for SH3 domains (occur at the position 19 in the alignment), are shown in red. They normally bind to a positively charged residue at P−3 in the binding peptides. Domain-specific determinant residues are shown in blue. For example, EPS8 prevents the interaction in case of deletion of the Lys residue from the position 38. The MUSCLE program was used for the alignment [42].

domains contain a PxxP core. A position system has been induced where the first P of this motif is given P0, upstream positions are given P−1, P−2, and so on, and downstream positions are given P1, P2, and so on. SH3 binding motifs can be categorized into two main groups: canonical and non-canonical motifs.

- **Canonical motif:** The canonical motifs generally contain the ΦPxΦP core formed by two ΦP dipeptides. These motifs are further classified in two major groups: class I and class II. The consensus sequences for these two groups are denoted as +xΦPxΦP (class I) and ΦPxΦPx+ (class II). Here, x represents any naturally occurring amino acid, Φ represents a hydrophobic amino acid, and + represents a positively charged amino acid (normally Arg and Lys).

Structural studies of the SH3-peptide complexes with class I and class II motifs suggest that these two types of peptide ligands bind to an SH3 domain in opposite orientations by forming a left-handed helix, called the polyproline type II (PPII) [51, 59, 63]. SH3 motifs contain two XP dipeptides that reside in the core (XP-x-XP). These dipeptide units occupy two binding pockets of an SH3 domain by hydrophobic interactions. Previous studies reveled that the positively charged residues in the peptide sequence, such as Arg and Lys, play an important role in the interaction with the respective SH3 domain [64, 65]. Based on the characteristics of the binding site, the SH3 domains prefer either one or the other peptide motif. These motifs can be further classified into sub-groups depending on the tolerance for the substitution of the Lys residue with the Arg residue [58].

Although most SH3 domains bind to class I and/or class II motifs, a subset of SH3 domains have the ability to recognize non-canonical or atypical peptide motifs. There are several non-canonical motifs that have been identified previously. Here, we describe a few of them.

- **Non-canonical motif:**
  - **PxxDY:** SH3 domains from EPS8 family and the first SH3 domain from NCK1 protein have been reported to bind to a PxxDY motif [66, 67]. EPS8 protein and its SH3 domain play an essential role in mitogenetic signaling. Over expression of EPS8 increased epidermal growth factor (EGF) dependent transformation and mitogenic responsiveness to EGF [68, 69].
  - **ΦxxPxxP:** This motif reassembles the class I consensus, which contains a hydrophobic residue instead of a positively charged residue at position P$-3$. These motifs are known to interact with the SH3 domain of the ABL1 protein [70].
  - **RxxK:** First identified as the binding motif of STAM2 SH3 domain [71]. This motif was first found in deubiquitinating UBPY enzyme. It was also observed that the motif can mediate an interaction between the C-terminal SH3 domain of GRAB2 and SLP-62, although the disassociation constant for the interaction was significantly low, i.e., 1-10 nm [72, 73].
  - **RxxPxxxP:** This motif is similar to class I consensus, and found in cytoplasmic tail of the BK channel. The SH3 domain of CTTN are known to bind to this motif [74].
  - **PxxxPR:** This motif reassembles the class II consensus, and can interact with the domains that are known to interact with class II motifs [58]. The SH3 domain from CIN85/SH3KBP1 proteins is targeted by this motif [75].
  - **RKxxYxxY:** This tyrosine-based motif does not hold a proline residue, and identified in the adaptor protein SKAP55. This motif is targeted by the SH3 domain of ADAP/FYB protein [76].

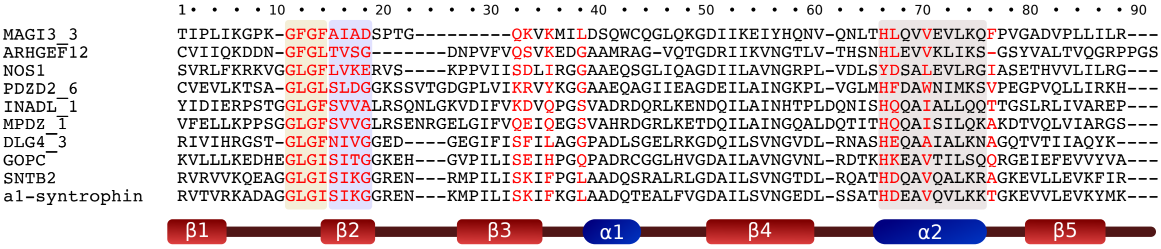
Other non-canonical motifs, such as PPxVxPY, RxxxxY, and RxxRxxS, have also been identified previously [77]. In our study, we used a peptide array data from [58] and reported the percentage of all canonical and non-canonical motifs bound by various SH3 domains (see Table A.2.1).

### 1.4.3. PDZ domains

Scaffold proteins are an important class of proteins that are indispensable for many key signaling pathways. The main role of these proteins is to assemble the multiple members of a signaling pathway into functional protein complexes to regulate signal transduction or localize signaling molecules to a specific compartment of the cell (e.g., cell membrane, nucleus, cytoplasm, mitochondria etc.) [78]. Scaffold proteins are composed of modular domains, which are mainly responsible for building the protein complexes by interacting with other proteins in the cellular space [79].

PDZ domains are one of the most promiscuous modular domains that are predominantly found in scaffold proteins in multi-cellular organisms, and play an important role in the establishment of cell polarity, cell signaling, protein trafficking etc. [80–82]. It has also been reported previously that PDZ domains take a pivotal role in several human diseases, such as schizophrenia, cystic fibrosis etc. [83]. In early 90's, the PDZ domain was discovered in three proteins, namely postsynaptic density protein-95 (PSD-95), disks large tumor suppressor (DLG1), and zonula occludens-1 (ZO-1) [84–86]. The domains were initially named as GLGF, since they have a repetitive motif (Gly-Leu-Gly-Phe) in their N-terminal sequences. Shortly after, the domains were renamed as DHR (Dlg homology region) domains; and finally, the name PDZ was derived from the acronym of these three proteins (i.e., PSD-95, DLG1, and ZO-1) for better reflection of the origin and distribution of the domain, which was then accepted by the scientific community [87]. Interestingly, PDZ domains are more abundant in multi-cellular organisms (e.g., human, mouse, plant, fly, worm etc.) than unicellular organisms (e.g., bacteria, archaea etc.), which may be due to the co-evolution of the PDZ domains [81]. For example, around 270 PDZ domains are found in more than 150 proteins in human proteome (see Section 4.4.2), but on the other hand yeast (*Saccharomyces cerevisiae*) proteome has three PDZ domains composed in only two proteins (`UniProtKB/-Swiss-Prot` database, release 2015-01 [19]).

PDZ domains can be observed in multiple copies in the proteins that are mainly found in cytoplasm. However, PDZ domains can also be seen in combination with other modular domains, such as SH3, PTB etc. Surprisingly, PDZ domains have never been seen with SH2 domains [82]. Based on the modular organization, PDZ domains are classified into three families: (i) in the first family, all proteins are entirely composed of PDZ domains (e.g., NHERF1). The number of PDZ domains can vary from 2 to more than 10; (ii) the second family proteins contain one or three PDZ domains, one SH3 domain, and one guanylate kinase-like (GK) domain. PDZ domains from this family are observed in membrane-associated guanylate kinases (MAGUKs), which include PSD-95, DLG1, and ZO-1; and (iii) the third family comprises proteins (e.g., MPDZ) that contain PDZ domains in combination with other protein domains (e.g., PH, WW, L27, LIM etc.) [82].



```
                 1• • • • •10 • • • •20 • • • •30 • • • •40 • • • •50 • • • •60 • • • •70 • • • •80 • • • •90
MAGI3_3          TIPLIKGPK-GFGFAIADSPTG---------QKVKMILDSQWCQGLQKGDIIKEIYHQNV-QNLTHLQVVEVLKQFPVGADVPLLILR---
ARHGEF12         CVIIQKDDN-GFGLTVSG-------DNPVFVQSVKEDGAAMRAG-VQTGDRIIKVNGTLV-THSNHLEVVKLIKS-GSYVALTVQGRPPGS
NOS1             SVRLFKRKVGGLGFLVKERVS----KPPVIISDLIRGGAAEQSGLIQAGDIILAVNGRPL-VDLSYDSALEVLRGIASETHVVLILRG---
PDZD2_6          CVEVLKTSA-GLGLSLDGGKSSVTGDGPLVIKRVYKGGAAEQAGIIEAGDEILAINGKPL-VGLMHFDAWNIMKSVPEGPVQLLIRKH---
INADL_1          YIDIERPSTGGLGFSVVALRSQNLGKVDIFVKDVQPGSVADRDQRLKENDQILAINHTPLDQNISHQQAIALLQQTTGSLRLIVAREP---
MPDZ_1           VFELLKPPSGGLGFSVVGLRSENRGELGIFVQEIQEGSVAHRDGRLKETDQILAINGQALDQTITHQQAISILQKAKDTVQLVIARGS---
DLG4_3           RIVIHRGST-GLGFNIVGGED----GEGIFISFILAGGPADLSGELRKGDQILSVNGVDL-RNASHEQAAIALKNAGQTVTIIAQYK----
GOPC             KVLLLKEDHEGLGISITGGKEH---GVPILISEIHPGQPADRCGGLHVGDAILAVNGVNL-RDTKHKEAVTILSQQRGEIEFEVVYVA---
SNTB2            RVRVVKQEAGGLGISIKGGREN---RMPILISKIFPGLAADQSRALRLGDAILSVNGTDL-RQATHDQAVQALKRAGKEVLLEVKFIR---
a1-syntrophin    RVTVRKADAGGLGISIKGGREN---KMPILISKIFKGLAADQTEALFVGDAILSVNGEDL-SSATHDEAVQVLKKTGKEVVLEVKYMK---
```

**Figure 1.7.:** Alignment and structural organization of human PDZ domains. Residues occur in the ligand binding pocket, which is formed by the second $\beta$ strand, the second $\alpha$ helix, and a GLGF loop, are represented with colored background. Residues in red colors are the main specificity determinants of PDZ domains [88]. The `MUSCLE` program was used for the alignment [42].

### Sequence and structure

PDZ domains are typically 80−90 amino acids in length, containing 5−6 $\beta$ strands and 2 $\alpha$ helices (see Figure 1.5.C). Figure 1.7 illustrates the sequence and secondary structural organization of PDZ domains. The overall sequence identity of PDZ domains is about 30%, and there are more than 300 structures available in `PDB` database [89]. The structure and function of a PDZ domain can be affected by an additional secondary structure. For example, PDZ3 from PSD-95 has an additional C-terminal $\alpha$ helix that significantly influences the function of the proteins [90]. It has been observed that the PDZ domains are highly resistance against extensive mutagenesis [91].

### Binding specificity

Most commonly, PDZ domains are known to bind hydrophobic C-terminal residues of target proteins [92]. The second $\beta$ strand, the second $\alpha$ helix, and a GLGF loop of the PDZ domain collectively form the binding pocket, which recognizes C-terminal peptides of their binding proteins [93, 94]. However, slight deviations of this concept have also been reported previously [95, 96]. For these C-terminal binding motifs, a global position system has been induced where C-terminal residue (i.e., last amino acid of a binding peptide) is given position P0 and going upstream residues are given position P−1, P−2, and so on. Last four residues of C-terminal motifs are known to be the most important to bind PDZ domains [94, 97], although there are some PDZ domains that are found to interact with residues up to position P−7 by an extended $\beta$2-$\beta$3 [98] loop or an extended $\alpha$2 [99]. Interestingly, some amino acids in C-terminal peptides are more preferred than other amino acids in specific position. Deletion or mutation of an important residue of a binding peptide can drastically reduce the interaction [92].

Due to high variability in sequences, PDZ domains are highly specific about their binding partners. In earlier studies, PDZ domains were grouped into three different classes based on their C-terminal binding motif structures: x[T/S]xΦ-COOH (class I motif), xΦxΦ-COOH (class II motif), and x[D/E]xΦ-COOH (class III motif), where x represents any natural amino acid and Φ represents hydrophobic amino acid [92, 100].

- **Class I motif (x[T/S]xΦ-COOH):** In this motif, serine (Ser) and threonine (Thr) are the most preferred amino acids in position P−2, and a hydrophobic residue is preferred in position P0. For example, second PDZ domain of PSD-95 interacts with ETDV-COOH and ESDV-COOH motifs of the Shaker-type $K^+$ channels and NMDA receptor, respectively [84, 101].
- **Class II motif (xΦxΦ-COOH):** Here, hydrophobic residues, such as Val, Tyr, Phe, Leu, Ile etc., are preferred in position P−1 and P−3. For example, PDZ domains from LAP2/ERBIN and CASK bind to DVPV-COOH and EFYA-COOH motifs of the ErbB2 and Syndecan proteins, respectively [102, 103].
- **Class III motif (x[D/E]xΦ-COOH):** This motif is comparatively rare. Here, aspartic acid (Asp) and glutamic acid (Glu) are preferred in position P−2, and a hydrophobic residue is favorable in position P0. For example, NOS1 PDZ domain recognizes VDSV-COOH motif of the Melatonin receptor [100].

Nevertheless, this classification system is an oversimplification, since it is known that every residue in the target peptide contributes to the binding specificity of respective PDZ domains. Although PDZ domains are most commonly found to interact with C-terminal tails of their target proteins, other interacting modes, such as interaction with internal peptides [104, 105], homo and/or hetero dimerization [106, 107], and binding with membrane phospholipid [108, 109], have also been previously detected.

## 1.5. Cellular and molecular function of the modular protein domains

### 1.5.1. Function of SH2 domains

Various biological processes are executed by SH2 domain mediated interaction. Different cellular and molecular functions of SH2 domains are described below.

#### Interaction with receptor tyrosine kinases (RTKs)

One of the main functions of SH2 domains is to bind activated receptor tyrosine kinases (RTKs) to regulate a series of biochemical pathways [12]. In human proteome, 90 tyrosine kinase genes were identified; among them 58 genes are responsible for encoding the receptor tyrosine kinase proteins [110]. An important RTK, i.e., platelet-derived growth factor receptor (PDGFR), recruits several SH2 domains upon its activation and transmits downstream signals that initiate various cellular processes, such as DNA synthesis, immediate-early-gene expression [111, 112]. The phosphorylated tyrosine residues in the cytoplasmic tail of PDGFR are targeted by various SH2 domains. However, the signaling strength differ qualitatively and quantitatively, since residues in the close vicinity of phosphotyrosine (pTyr) are the main determinants of the binding specificity of a SH2 domain [112].

In signaling process, several adaptor proteins, such as GRB2, SHC, and NCK, are also recruited to activated RTKs. For example, GRB2 protein uses its SH2 domain to bind RTKs and uses its two SH3 domains for further interacting with SOS, a Ras guanine-nucleotide exchange factor (Ras GEF), which facilitates the downstream signaling for mitogen-activated protein kinase (MAPK) pathway [112, 113].

Previous research revealed that the lack of epidermal growth factor receptor (EGFR) signaling in human can cause several neurodegenerative diseases, such as Alzheimer's disease and multiple sclerosis [114]. Additionally, it has also been observed that the mutation of two important tyrosine residues of hepatocyte growth factor receptor (HGFR), which mainly interacts with PI3K, GRB2, SHC, and SRC SH2 domains, can cause an embryonic lethal phenotype in mice [112, 115], and altering these tyrosine residues in a way by which only a subset of SH2 domains can bind will cause complex cell-specific effects [112, 116].

#### Interaction with cytoplasmic protein tyrosine kinases

The SRC family kinases, e.g., YES, FYN, FGR etc., are a family of non-receptor kinases that facilitate various signaling activities by interacting many cellular and nuclear proteins. Previous studies showed that the SH2 domains from SRC family proteins have two main functions: (i) keeping the kinases in an inactive state. For example, SH2 domain from SRC protein in the chicken interacts with its own C-terminal pTyr residue (pY527) to maintain the inactive state; and (ii) play an important role in processive phosphorylation and substrate targeting [112].

#### As a negative regulator

SH2 domains are also known to play as a negative regulator in cellular signaling. For example, SH2 domain from RasGAP is a negative regulator of Ras. By catalyzing, it converts an active GTP-bound form to an inactive GTP-bound form of Ras and thus plays a crucial role to control the Ras signaling [112, 113]. Another example of negative regulator is c-CBL SH2 domain, which interacts with phosphorylated EGFR and degrades the activator receptor [112].

#### As an *"allosteric switch"*

Currently, there are 10 proteins available in `UniProtKB/Swiss-Prot` [19] database, which contain two tandem SH2 domains (N-terminal and C-terminal). They are known to interact with multiple phosphorylated ligands. For example, SH2 domains from the ZAP-70 protein bind to the immunoreceptor tyrosine-based activation motifs (ITAMs) [112]. An appropriate example that explains the function of the tandem SH2 domains is the PTPN11 mediated interactions. In one structural conformation of PTPN11, the N-terminal SH2 domain binds to its protein tyrosine phosphatase (PTP) domain and thus maintains the inactive state of the phosphatase. In a different conformation of PTPN11, when the N-terminal SH2 domain binds to a phosphotyrosine peptide, the phosphatase becomes in an active state [112, 117]. Therefore, in this case, the N-terminal SH2 domain of PTPN11 acts as an *"allosteric switch"* [112].

#### Association with human diseases

Previous studies showed that the mutations in some SH2 domains are involved in several human diseases. For example, the X-linked lymphoproliferative (XLP) syndrome can be occurred by the disruption or a point mutation of SH2D1A/SAP SH2 domain [27]. The mutation in the N-terminal SH2 domain of the PTPN11 protein can cause the Noonan syndrome, which is an autosomal dominant congenital disorder [28]. It has also been observed that the X-linked $\alpha$-gammaglobulinemia can be caused by a point mutation in the SH2 domain of Bruton's tyrosine kinase (BTK) [29, 118].

Importantly, in the last decade, SH2 domains have become one of the important candidates for the drug discovery. SH2 domains from ZAP-70, SRC, GRB2, LCK, and PI3K have been found as the potential candidates that could be used for several disease treatments, including cardiovascular disease, osteoporosis, immune system disorder, and cancer [112, 119].

### 1.5.2. Function of SH3 domains

For a long time after discovery, the function of SH3 domains in eukaryotes was elusive. However, subsequent research showed the significant contribution of SH3 domains in a diverse range of signaling processes, such as regulation of enzymes, altering the subcellular localization, and controlling the assembly of large protein complexes [49, 120–122]. The cellular and molecular functions of SH3 domains are discussed below.

#### Intramolecular interaction

An appropriate example of the SH3 domain mediated intramolecular interaction can be found in the proteins from the SRC family. It is known that these proteins maintain their inactive state by an intramolecular interaction where the C-terminal phosphotyrosine peptide interacts with its own SH2 domain [49]. However, the structural studies of SRC and HCK revealed that SH3 domain binds to a linker region, which contains a PPII helical conformation, between the kinase and SH2 domain in the same protein [49, 123, 124]. These observations have further revealed that the SH3-linker intramolecular interactions are essential for holding the cytoplasmic enzymes in their inactive conformation.

#### Effect of phosphorylation

Phosphorylation event takes a critical role in SH3 domain mediated interactions. For example, the interaction between the SH3 domain of cytoskeletal-associated protein (PST-PIP) and Wiskott-Aldrich syndrome protein (WASP) is inhibited by phosphorylation of

PxxP segment of WASP [125]. Phosphorylation of serine/threonine residues in proline-rich C-terminal region of SOS protein inhibits the interaction with GRB2 protein [126]. Other effects of phosphorylation that regulate SH3 mediated interaction have also been observed [127, 128].

Interestingly, phosphorylation events can serve as a *"molecular switch"* for SH3 and SH2 mediated interactions. The CD3epsilon contains an SH3 domain binding motif, i.e., PxxDY, which also have a tyrosine (Y166) residue and thus upon phosphorylation this motif should also be capable to bind an SH2 domain. Kesti *et al.* showed Eps8L1 and NCK-N SH3 domains bind this CD3epsilon containing motif, but phosphorylation of the tyrosine (Y166) residue enables ZAP-70 SH2 interaction while diminishing SH3 interaction [66]. Here, phosphorylation of the tyrosine residue (Y166) acts as a *"molecular switch"*. Other such examples of the *"molecular switch"* have also been reported [129–131].

#### Maintaining the assembly of large multi-protein complexes

The proteins from membrane-associated guanylate kinase (MAGUK) family are composed of 1−3 PDZ domains, an SH3 domain, and a guanylate kinase-like (GK) domain. The MAGUKs facilitate SH3 domain mediated intramolecular and intermolecular interactions and play an important role in assembly of large multi-protein complexes at specific membrane regions, such as cell-cell junctions, synapses, and neuromuscular junctions [49, 132]. Previous researches showed that SH3 domains of numerous MAGUKs, including PSD-95, DLG, CASK, and p55, interact with their own GK domains [49, 120, 133]. The intermolecular interactions occur when an SH3 domain of a MAGUK binds to a GK domain of a second MAGUK. In *Drosophila sp.*, it has been observed that mutation in SH3 domain in a MAGUK protein, i.e., DLG, causes malignant transformation and epithelial overgrowth [49, 121].

#### Controlling actin

The Las17p/Bee1P, a WASP homologue, is an important protein for the assembly of cortical actin cytoskeleton and endocytosis in yeast [134, 135]. These WASP family proteins are often targeted by the SH3 domains from myosin proteins. Mutation of the type I myosines, i.e., Myo3p and Myo5p, results severe defects in the actin polarization [122, 136, 137]. Furthermore, Myo3p and Myo5p are also found to bind the verprolin, a protein-rich protein, and play a critical role in the actin cytoskeleton organization [122, 138].

### 1.5.3. Function of PDZ domains

Initially, it was known that PDZ domains are only responsible for assembling the signaling molecules to regulate signal transduction, but gradually it got clear that they do more than it was thought. Different cellular and molecular functions of PDZ domains are described below.

### Regulation of PDZ domains

For understanding the diverse functionality of PDZ domains, one has to acquire knowledge about the regulatory mechanisms of PDZ domain mediated interactions. The post-translational modification (PTM), allosteric changes, and autoinhibition have been identified as the reasons for the regulation of PDZ domain mediated interactions [139].

Phosphorylation is a prime example of post-translational modification that takes an important role in the regulation of PDZ domains. A serine phosphorylation of the NR2B subunit of NMDA receptor (S1480) disrupts its interaction with the PSD-95 PDZ domain, which eventually reduces the surface expression of the NR2B in neurons [140]. Interestingly, is has been observed that phosphorylation of PDZ domain itself can also negatively regulate the PDZ mediated interactions. A serine phosphorylation of PDS-95 PDZ domain (S73) negatively regulates the spine growth and synaptic plasticity [141]. More recently, Akiva *et al.* showed phosphorylation can serve as a *"molecular switch"* or *"specificity switch"* for PDZ mediated interaction [142]. They showed that PDZ domains have inverse affinities to the phosphorylated and non-phosphorylated peptides. This study indicates that in a normal state, some potential motifs may interact with their specific PDZ domains, but upon phosphorylation, they reject those specific PDZ domains and bind to different PDZ domains [142].

PDZ domains are functionally very dynamic, which can be regulated by the allosteric behavior of PDZ containing proteins. Van der Brek *et al.* have shown that the intramolecular PDZ-PDZ interaction allosterically modulates the binding preferences of PDZ domains. They found the binding specificity of PTP-BL PDZ2 can be modulated by the presence of PDZ1 [143]. In this case, the PDZ1 binds to the surface on PDZ2, which is opposite to the peptide binding groove. A recent genome-wide study revealed that 40% of PDZ domains have lipid membrane affinity and act as dual-specificity modules, which regulate protein interactions at the membrane [144].

In a specific structural conformation, intramolecular PDZ mediated interactions can cause autoinhibition. For example, if a C-terminal tail of a protein interacts with its own PDZ domain then it is no more accessible for further binding and therefore adopt the auto-inhibitory conformation [139]. Several examples for autoinhibition of PDZ domains have been reported for NHERF1, X11 $\alpha$, and tamalin/GRP1-associated scaffold protein [96, 145, 146].

### Adaptor for receptor tyrosine kinases (RTKs)

It has been previously observed that the PDZ domains can play an important role in localization of the signaling molecules, such as receptor tyrosine kinases and glutamate receptors [78, 82]. For example, in *Caenorhabditis elegans*, the Lin-7-Lin-2-Lin-10 PDZ protein complex interacts with the Let-23, an epidermal growth factor receptor homologue and can mislocalize it so that it cannot access its binding ligand, Lin-3 [82, 147]. Furthermore, autophosphorylation of the platelet-derived growth factor receptor (PDGFR) is also caused by interacting with a PDZ domain from NHERF [148]. Basically, PDZ domains provide a versatile option to the RTKs for specifying their functions [82].

### Maintaining epithelial polarity

PDZ domain containing proteins are often localized near cell membrane of a polarized cell and thus take a central role to maintain the epithelial polarity of the cell [82]. Bilder *et al.* showed that the Dlg, lgl, and Scrib proteins are required for regulating the epithelial cell polarity [149].

### Roles in synaptic communications and protein networks

PDZ domains play a crucial role in synaptic communication. In mammalian central nervous system, the neurotransmitter receptors, such as glutamate receptors, activate various signaling pathways by interacting with PDZ domain containing proteins [82]. For example, N-methyl-D-aspartate (NMDA) receptors bind to the PDZ domains from PSD-95 and play a key role in synaptic plasticity [101]. Furthermore, it has been observed that the PDZ domain containing synaptic proteins are important to build a large protein network [82].

### Association with diseases

PDZ domains have a large association with numerous diseases. A PDZ domain containing protein, namely Shroom, is involved in several diseases, such as spina bifida, cleft palate, acrania etc. [82, 150]. In mice, a mutation in the Dlg genes causes craniofacial dysmorphogenesis with cleft palate [151]. Disruption of Scrib protein causes a severe neural tube defect, and evolutionary high conservation of this protein also suggests it may have an effect in tumorigenesis [82, 152, 153]. Additionally, previous study suggested that the disruption of PDZ domains has a deep impact on various signaling pathways that are found in cancer [82].

Chapter 2

# Overview of existing methods and their limitations

## 2.1. Overview

Thousands of modular protein domains that recognize linear peptides are spread across the eukaryotic genomes. Accomplishing large-scale data about binding specificities for all peptide recognition domains is a very challenging task. However, many high-throughput methods have been successfully introduced to address the binding specificity of some peptide recognition modules. The enormous data generated by these high-throughput experiments have become invaluable to build powerful computational models. In this chapter, we first discuss about the important high-throughput techniques that have been widely used for identifying modular domain-peptide interactions along with their limitations. We then discuss various computational methods that have been developed for predicting modular domain mediated interactions. Majority of these methods train their prediction models using the data generated by high-throughput techniques. Finally, we highlight the drawbacks of existing methods that can severely affect the prediction accuracy. Some parts of the publication [P4] are presented in this chapter.

## 2.2. High-throughput techniques

Over the years several experimental approaches have been employed to identify *in vitro* binding specificity of modular protein domains. High-throughput (HTP) analysis of modular protein domains using peptide arrays (SPOT, OPAL etc.), microarrays, phage display, and other HTP techniques are invaluable to understand the underlying nature of their binding specificity, and thus take an important role to define the potential domain-peptide interactions. However, each HTP technique has particular limitations and pitfalls. See Table 2.1 for the outline of pros and cons of established HTP techniques.

In the following sections, we will describe some of the major techniques that have been widely used for describing the specificity of modular protein domains.

### 2.2.1. High density peptide arrays

Peptide array technology is a powerful tool that has been successfully used for defining binding specificity of modular protein domains, screening for cellular interaction partners, and developing selective protein interaction domains inhibitors [43, 58, 155–157].

In 1983, Roland Frank described the initial concept of simultaneous synthesis of multiple components on a solid support [158]. Almost a decade later, two techniques of chemically synthesizing peptide arrays were developed: the SPOT synthesis technique was pioneered by Frank *et al.* [159] and light-detected, spatially addressable parallel chemical synthesis was described by Fodor *et al.* [160]. The SPOT synthesis technique was widely accepted by the scientific community, since it is very simple and extremely robust technique for parallel synthesis of thousands of peptides and subsequently screen them on a solid surface. Over the years the technique has been reviewed several times and has gradually becomes an important tool in biology, particularly in molecular immunology [161]. In SPOT synthesis, when small droplets of activated amino acids are spotted onto the planar surface of a porous cellulose membrane (see Figure 2.1.A), the droplets are absorbed and form circular spots, and

**Table 2.1.:** The table contains common high-throughput techniques that determine the binding specificity of modular protein domains, and their pros and cons. The table is adapted with permission from *John Wiley and Sons: Proteomics* [154] (license number: 3558260224349).

| Methods | Library size | Quantitative | Pros | Cons |
|---|---|---|---|---|
| Peptide array | 10-1000s | Semi-quantitative | Unnatural and modified amino acids, PTMs, produces negative binding data, and easy to generate different libraries | Biased libraries, high cost of materials |
| Protein microarray | 10-100s | Quantitative | Quantitative | Protein stability, limitation in number of peptides |
| Phage display | 1 x 10ˆ10 | Not quantitative | Random peptides, low costs for production | Only natural amino acids, no PTMs, high cost in DNA sequencing |

therefore efficiently generate an open reactor for synthesis of cellulose-bound peptides [161]. Automation of this technique is also possible with a multiple synthesizer (Intavis AG, Köln, Germany) in analytical and preparative mode that enable parallel synthesis of upto 6000 and 1000 cellulose membrane-bound peptides, respectively [155].

A variation of synthetic library namely, oriented peptide array library (OPAL) was developed by Rodriguez *et al.* in 2004 [156]. The OPAL approach integrates both oriented peptide libraries and array technologies that allow to synthesize hundreds of pools of oriented peptide libraries and are arranged as scan arrays. OPAL requires less knowledge about binding preference of domains but generally works best when at least one important position of the binding peptide is fixed [156]. For example, SH2 domains recognize phosphorylated tyrosine (pTyr) residue with high affinity and hence the position of pTyr residue in the array can be fixed and a randomized mixer of amino acids can be used in other positions. Some domains target their peptides with relatively lower affinity (e.g., SH3, PDZ etc.) and in this case, at least two residue positions need to be fixed [154].

More recently, Tinti *et al.* has proposed a peptide chip technique, which is also a variation of SPOT synthesis, to study SH2-peptide interactions. This method is capable of synthesizing a much higher number of peptides (several thousands) in a single experiment and screen them onto aldehyde-modified glass surface [162].

Peptide array technique has several advantages: (i) it allows the use of unnatural and modified amino acids as building blocks in the peptide synthesis, which is very useful to study the interactions that depend on post-translational modification (PTM) of the binding peptides, (ii) a modified SPOT technique can synthesize membrane-bound peptides with free C-termini, which allows to study the PDZ mediated interactions [163], and (iii) along with the capability of producing binding interactions, it also has the ability to identify weak binding (if weak signal is detected in the array) and non binding (if no signal is detected in the array) interactions. However, peptide array technique has some disadvantages as well: (i) it needs some degree of priori knowledge of binding specificity of the modular domains, (ii) the efficiency of the peptide synthesis on a solid surface is not uniform and may be difficult to assess, since the peptides on the array do not undergo purification. Hence, the results could be affected by false negative interactions [154].

### 2.2.2. Protein microarrays

In protein microarray, the modular proteins are typically immobilized onto a solid surface, such as a modified glass microscope slide, and probed with enzyme-labeled peptides [164]. The flexible detection methods (fluorescence-based and enzymatic detection) and attachment method are probably the key components behind the success of protein microarrays. Protein microarray has become a versatile tool, which is suitable for large-scale analysis.

For determining the specificity of the modular protein domains, protein microarray has been successfully used in the last few years (see Figure 2.1.B). MacBeath and co-workers developed protein interaction maps using protein microarrays to investigate SH2/PTB pep-

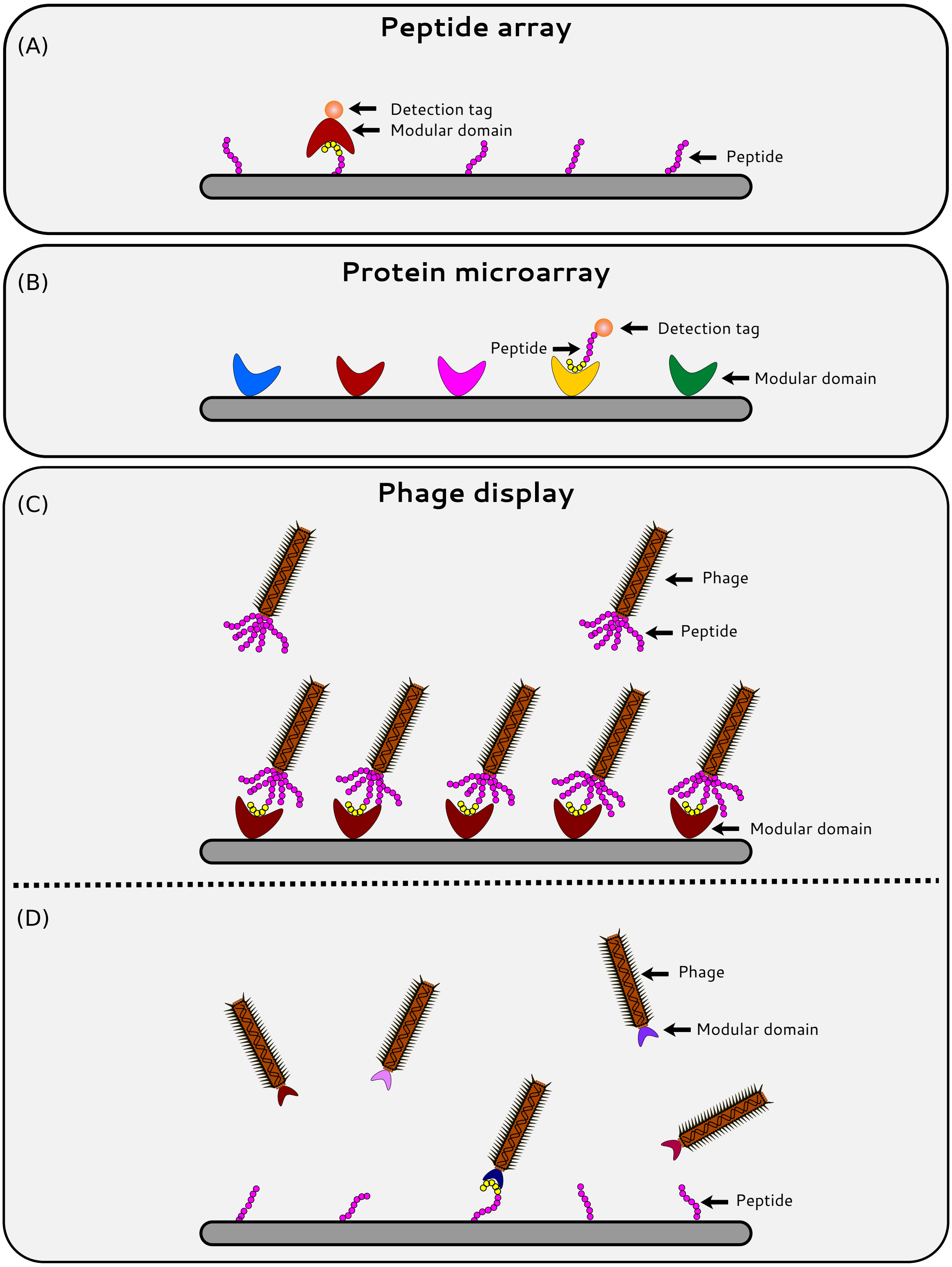


**Figure 2.1.:** High-throughput techniques to determine the binding specificities of modular protein domains. (A) Peptide array: a diverse set of peptides are immobilized on a solid surface and probed with an interacting domain. The binding is detected by fluorescence or antibody based system. (B) Protein microarray: a set of soluble domains are immobilized on a solid surface and probed with a labeled peptide. A fluorescence-based system is used for the interaction detection. (C+D) Phage display: (C) a diverse set of random peptides are expressed on the bacteriophage coat and incubated with immobilized soluble domains, and (D) in a reverse process, a diverse set of domains are expressed on the bacteriophage coat and screened to bind peptides. High affinity interactions are detected using sequencing of phage DNA.

tide interactions [44, 165]. Stiffler *et al.* fabricated another protein microarray to identify the PDZ peptide interactions [164]. These methods calculate the affinity of domain peptide interactions. To assess the accuracy, an apparent equilibrium disassociation constant ($K_D$) was used. More recently, a cellulose peptide conjugate microarray (CPCMA) has also been developed to quantify the specificity of SH2 domains [166].

However, this method also has some disadvantages: (i) the required orientation and conformation for the peptide binding might be disrupted by immobilization method of the recombinant proteins. Thus, optimization with meaningful evaluation is required for the immobilization of a given protein [167], (ii) uses limited number of peptides to identify the modular domain peptide interactions, and (iii) the protein microarray techniques are often affected by a high rate of false positive and false negative.

### 2.2.3. Phage display

In 1985, Smith and co-workers first described the phage display technique [168]. Over the years, this technique has been improved and has gradually become one of the most powerful and conventional tools for proteomics screening. Over the years, several variations of the phage display have been applied to study protein-protein interaction [169, 170]. In this technique, the DNA are expressed as a protein and subsequently fused with phage coat protein to make a hybrid fusion protein. Short peptides are expressed as N-terminal fusions, however, C-terminal fusion is also possible [171]. More than 10 billion random peptides can be screened by the phage display libraries [97]. High affinity scores can be detected by the biopanning process that generally includes four steps: (i) phage display library preparation, (ii) capturing or panning, (iii) washing, and (iv) elution [172]. Phage libraries that bind to the domains can be easily isolated and sequenced to determine the binding specificity of modular protein domains. Beside the principle to express the peptide sequences, modular domains can also be expressed on the phage surface (see Figure 2.1.C and 2.1.D). Karkkainen *et al.* expressed SH3 domains on phage surface and showed that these domains can bind their target peptides with a much higher affinity than previously reported [54]. Phage display technique has also been used to study the mutational tolerance and the *in vitro* evaluation of a modular domain [91]. Phage display has been proven as a powerful method to investigate modular domain mediated interactions. Tonikian *et al.* efficiently employed phage display technique to determine the binding specificity of PDZ and SH3 domains [97, 173]. A combined strategy, which includes phage display and large scale yeast two-hybrid methods has also been used to identify the relevant binding partners of SH3 domains from yeast proteome [77]. Main advantages of this technique are: (i) a large set of chemically diverse peptide ligands can be produced efficiently [97] and (ii) inexpensive production costs. However, this method also has some pitfalls: (i) unlike SPOT synthesis, the phage display technique does not include unnatural amino acids in the phage library, which makes it difficult to study the PTMs and (ii) it is a bit expensive for DNA sequencing [154].

## 2.3. Existing computational methods

While the enormous amount of data generated by these aforementioned high-throughput techniques have become very important to describe specificity landscapes of different PRMs, however, obtaining large sets of such data for all peptide recognition domains is unfeasible due to some experiment stringencies, such as solubility, expression, common cloning etc. [174]. Furthermore, these kind of data also have some severe caveats. For example, data may be rich only for certain domains while it is scarce or completely missing for other domains. Thus, developing powerful computational methods, which infer binding specificity from a limited set of experimental evidences is indispensable. Several computational methods that are based on bioinformatics, statistics, machine learning approaches have been developed over the years. In this section, we will describe some important computational models that have been employed to predict binding partners of modular protein domains.

### 2.3.1. PWM-based methods

The interactions mediated by modular protein domains are highly selective, since these domains bind to their physiological partners that contain specific sequence patterns. For example, WW and Class I SH3 domains bind to peptide sequences mainly containing PPxY and [R/K]xxPxxP sequence motifs, respectively. An alignment of these peptide sequences for a specific domain could produce consensus sequences that can be used for describing the binding specificity of that domain. Few databases containing large collection of such consensus sequences for several modular domains have been reported recently [175, 176]. However, in many cases these consensus sequences are not enough to describe the specificity of a certain domain, as some less preferable residues in some position are often ignored. It is known that most of the SH3 domains bind to peptides that contain PxxP core motifs, however, previous research showed several unconventional motifs can also be targeted by the SH3 domains with slightly weaker affinity [66, 67, 74, 75]. For example, SH3 domain from CIN85 protein targets PxxxPR motif, which is slightly different than class II consensus but missing a PxxP core motif (see Section 1.4.2 for more details) [75]. Moreover, some amino acids are weakly preferred by the positions denoted with x in the consensus sequences. These preferences are often crucial to explain the binding specificity of different domains from the same domain family. To overcome these limitations, Stormo and his co-workers introduced Position Weight Matrices (PWMs) or Position Specific Scoring Matrices (PSSMs) in early 1980s [177, 178].

The main concept of this method is to compute probability scores for each amino acid at each ligand position. For a given domain, a verified set of known ligands need to be aligned. The probability can then be computed as the frequency of each amino acid at each position and subsequently a probability matrix can be constructed. Finally, a PWM score of a given peptide can be computed by multiplying the probabilities of different residues in different positions. The amino acid composition of the binding peptides can be visualized

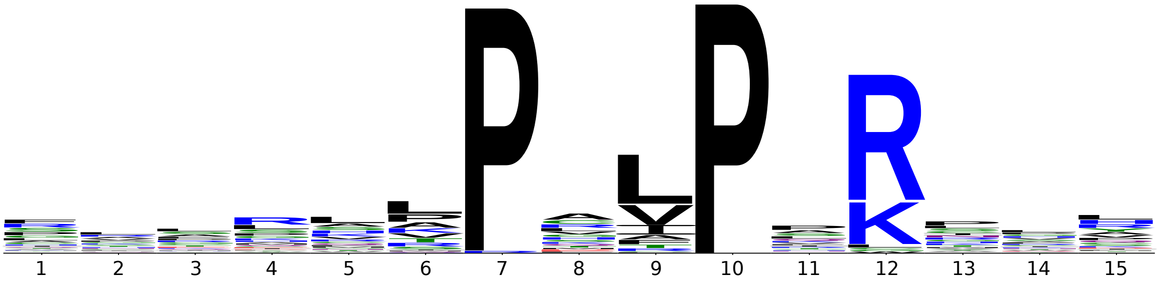


**Figure 2.2.:** This figure represents a sequence logo of class II motif containing peptides that are bound by SRC SH3 domain. Here, it is clearly observed that position 7, 10, and 12 have preferences for proline, proline, and arginine/lysine, respectively, and no amino acid preferences are observed in other positions. `WebLogo` was used for constructing the sequence logo [180].

using sequence logos [179]. The height of a letter at a certain position in the sequence logo is proportional to the frequency of the corresponding amino acid. Fully specific positions can easily be distinguished from a random position. For example, a sequence logo of class II peptides that are targeted by SRC SH3 domain is depicted in Figure 2.2; where position 7, 10, and 12 have preferences for specific amino acids but other positions do not have any specific amino acid preferences.

Over the last several years, many computational tools based on PWMs have been developed to address the specificity of modular protein domains. Most important tools are discussed in the following sections.

### Single PWM-based method

In single PWM-based methods, the specificity of a domain is represented by a single PWM. One of the most popular tools, `Scansite`, was introduced by Yaffe and co-workers in 2003 [181]. This tool is typically based on PWMs derived from chemically synthesized peptide array libraries and phage display experiments [182]. The tool has been widely used for predicting the peptides that are recognized by modular protein domains, phosphorylated by protein kinases (Ser/Thr/Tyr kinases), and mediate protein or phospholipid ligand interactions. Currently, 65 sequence motifs are available and each sequence motif is represented as a PWM. These motifs characterize binding specificities of many families of Ser/Thr/Tyr kinases, SH2, PTB, SH3, PDZ, and 14-3-3 domains [181].

Few years later, another single PWM-based method, scoring matrix-assisted ligand identification (`SMALI`), was developed by Li and colleagues to predict SH2 and PTB domain mediated interactions [183]. SH2 and PTB domains typically recognize phosphorylated tyrosine (pTyr) residue to perform their biological functions. `SMALI` PWMs were constructed from screening oriented peptide array libraries (OPAL)[39]. `SMALI` is similar to `Scansite` but it has a few additional advantages: (i) `SMALI` offers PWMs for 76 SH2 domains in contrast to 14 PWM models offered by `Scansite`, (ii) `SMALI` uses selectivity information for six positions (i.e., -2 to +4 amino acids with pTyr in 0th position) of a peptide to construct the PWMs; where most of the `Scansite` PWMs for SH2 domains were constructed using selectivity information for three positions (i.e., +1 to +3 amino acids with pTyr in 0th position), and (iii) to achieve more physiologically relevant interactions, SMALI incorporates additional filters, such as phosphorylated peptides, signal transduction, and subcellular localization of domain containing and binding proteins. A SMALI score is also normalized by an experimentally determined cut-off value. To determine the cut-off value, the author validated the top SMALI predictions for BRDG1 and GRB2 SH2 domains. They achieved best SMALI scores (cut-off values), based on F-measure, that separate top 3.5% and 5.5% as confident interactions for BRDG1 and GRB2, respectively. For other domains, they considered the SMALI scores, which separate top 4.5% (average percentage of BRDG1 and GRB2) interactions (see PSSM-based SMALI model in Section 3.2.2). Note that every domain has a different cut-off value. Tonikian *et al.* employed a single PWM-based model derived from phage display experiments to accurately map binding specificity for hundreds of PDZ domains in the human and worm proteome [97].

A multi-domain selectivity model (MDSM) derived from a protein microarray experiment has been proposed by MacBeath and co-workers [164]. This model is a variation of a PWM, which was designed to describe the difference in selectivity of many members of PDZ domain family. Other single PWM-based methods have also been developed to predict the modular domain mediated interactions [184–186].

Initially, it was assumed that residues in the peptides are independently responsible for the interaction. More clearly, the presence of an amino acid at a certain position is not influenced by the presence of other amino acid at other position in the binding peptides. This also implies that single PWM approaches cannot differentiate between peptide classes. However, based on this assumption of linearity, over the past years, the aforementioned single PWM-based approaches have been developed to build powerful computational models for many high-throughput data, specifically generated by oriented peptide array libraries [187].

### Linearity issues: a review on positional dependency problem

In 2010, Liu *et al.* showed that SH2 domains have distinct selectivity on their binding peptides. The underlying truth of the binding peptides is that they are composed of permissive and non-permissive amino acids, where permissive amino acids promote the interaction and non-permissive amino acids inhibit the interaction, and thus allows us to understand the subtle differences in peptide ligands [43]. This observation prompted that the residues in the close vicinity of phosphotyrosine are highly predictive for SH2 domain mediated interactions (see Figure 2.3 for details). It is known that the CRK SH2 domain binds peptides where amino acid Leu or Pro is present in position +3, however, presence of other amino acids in other positions can prohibit or even diminish the interaction. For example, basic residues (i.e., His and Arg) are disfavored in position +1 and +2, Ala is disfavored in position +1 and Pro is prohibited in position +1 and +2. Similarly, GRB2 SH2 domain interacts peptides with an Asn residue in position +2. Additionally, it also favors Glu in position +1, +3, and +4 but exhibits a prohibition against Arg and Asp in position +1, +3,

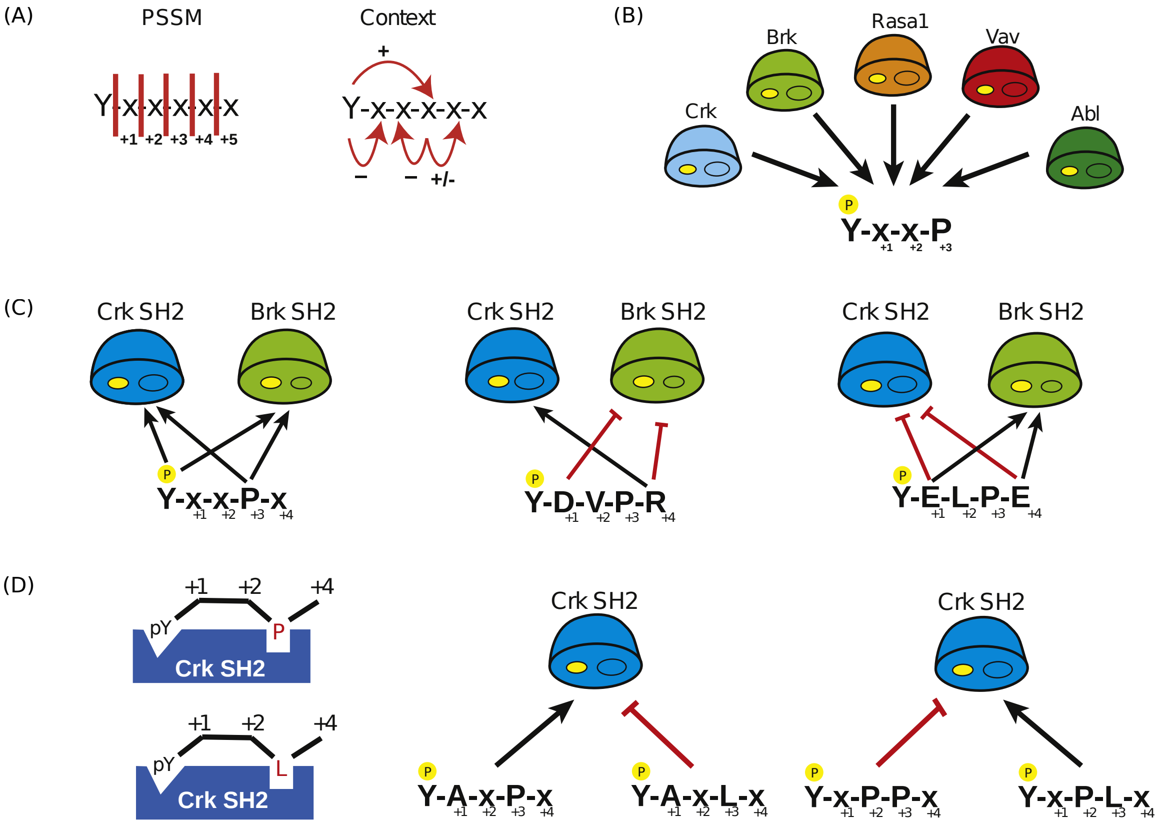


**Figure 2.3.:** Contextual binding specificity of the SH2 domains. (A) Left panel: PWM or PSSM based approaches consider each amino acid independently for each position. Thus, do not allow the positional dependency among the amino acids. Right panel: contextual specificity of the SH2 domains is shown, which depends on permissive and non-permissive amino acids in the peptide ligands and therefore indicates positional dependency. (B) SH2 domains from CRK, BRK, RASA1, VAV, and ABL families prefer a basic peptide motif, pY-xxP. However, sequence context and non-permissive residues of this motif take a vital role for discriminating the binding specificity of these domains. (C) Permissive and non-permissive amino acids. Left panel: general motif preference of CRK and BRK SH2 domains where they bind to a Pro at +3. Middle panel: while an Asp at +1 and an Arg at +4 are permissive and favored by CRK, they are non-permissive and disfavored by BRK. Right panel: Two Glu at +1 and +3 are favored by the BRK, whereas they are disfavored by CRK. (D) Contextual sequences of CRK mediated interactions. Left panel: basic binding specificity of CRK SH2 domain where it binds to a peptide, which contains a Pro or Leu at +3. Middle panel: an Ala at +1 with a Pro at +3 is favored, but an Ala at +1 with a Leu at +3 is disfavored by CRK. Right panel: a Pro at +2 with a Leu at +3 is favored, but a Pro at +2 with a Pro at +3 is disfavored by CRK. The phosphorylated Tyr is indicated by a P symbol with yellow background. The figure is adapted with permission from *Elsevier: FEBS Letters* [36] (license number: 3572171346505).

and +4, and rejects Lys at position -1, +1, +3, and +4 [43]. By looking at the examples mentioned above, it is clear that the selectivity of SH2-peptide interactions are not depending on physio-chemical property of the amino acids, as GRB2 SH2 domain favors Glu but rejects Asp (both are acidic amino acids) in the position +1, +3, and +4. This kind of highly selective nature of SH2 domains creates more difficulty to differentiate between the ligands having minor differences in their physio-chemical properties along with the struc-

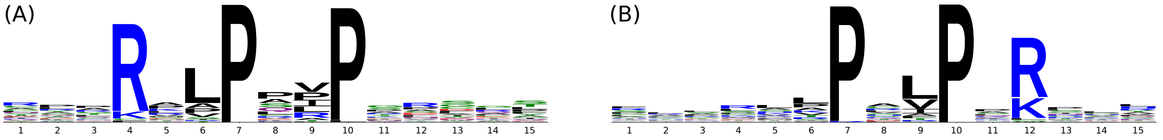


**Figure 2.4.:** (A+B) Multiple specificity of SRC SH3 domain. This figure represents the sequence logos of class I (A) and class II (B) motif containing peptides that are bound by SRC SH3 domain. `WebLogo` was used for constructing the sequence logo [180].

ture [43]. PWM-based approaches, such as `SMALI` and `Scansite` were appear to perform a good job for the prediction of binding ligand based on permissive residues. However, they were failed to recognize the importance of non-permissive residues [43], since they only rely on positive interaction data (generative approach).

More interestingly, it has been observed that there are positional dependency between non-permissive and permissive residues in the peptide ligands that interact with SH2 domains [36]. For example, the role (whether as permissive or non-permissive) of an Ala in position +1 or a Pro in position +2 depends on the occurrence of a permissive residue (Pro or Leu) in position +3 (see Figure 2.3.D). Highly significant dependencies between the ligand positions in other modular domains, such as SH3, PDZ, WW, Chromo, and 14-3-3 domains were also found [188–190]. For example, first PDZ domain from human DLG1 protein binds with peptides where Ile in the position -1 (C-terminal residue is given position 0) always appear with Trp or Leu at position 0, but is never found with Val at the same position. Thus, the local sequence context and the subtle dependency between the amino acids are highly important to define the binding specificity of a modular protein domain [36, 43, 188].

Recently, several studies have shown that dependencies between different ligand positions take an important role in binding specificity of modular domains [43, 188, 191]. These kind of positional dependencies between the amino acids in the binding peptides are completely ignored by the single PWM-based methods. Moreover, single PWM-based approaches assume that the domains are bound by the single class of peptides, i.e., all domains follow the same binding mode and hence are unable to describe the multiple specificity binding mode of modular domains. For instance, peptides that bind to SRC SH3 domain can be classified into two major classes: (i) class I consists [R/K]xxPxxP motifs and (ii) class II consists PxxPx[R/K] motifs (see Figure 2.4).

### Multiple PWM-based method

To circumvent the limitation of single PWM approach, Gfeller and colleagues proposed a mixture model that contains multiple PWMs to define the multiple specificity of a modular domain and showed that the multiple specificity model can predict protein interactions more accurately than single PWM-based model [188, 192]. They also showed that the same domain can bind to a set of peptides with distinct sequence patterns and these sequence patterns can be identified by a set of small number of peptide clusters where each cluster represents a sequence pattern. Recently, they have introduced a tool, MUltiple Specificity

```
---ACFRPPPLLPIRPCC-----
---AEMRARLLPPLPGLE-----
---AFKPPVPPRPQAKVP-----
---AGALARPKVPSRNRV-----
----AKQPPVPPPRKKRIS----
---AKTRPLPPLPPRLEC-----
---CKKLSPPPLPPRASI-----
----CKSLPLPPPRPPLLS----
---CKYRYLPERPHLRRL-----
---CLRPAPPLRPSAALC-----
----CYARRLPPRPTRSPA----
----FRPLLPRRPPGCGQH----
---FSRSLKPVLPHKVAH-----
GNLRCLPLPPRPPAT--------
---GPARGLPSLPLAGFS-----
----HPGGPVPPPRLLHLC----
---LVLPVVPLLPTRLSR-----
----LVPFPPPPPRTPLLW----
---PPARALPFPPPWAMQ-----
---TFRPLPPPPPPPHAC-----
```

**Figure 2.5:** This figure illustrates an alignment of SRC SH3 domain binding peptides derived from a high density peptide array screening experiment [58]. The alignment is generated by MUSI, which uses the MAFFT algorithm. Internal gaps were eliminated by increasing gap opening penalty [192, 195]. To make it simple, a subset of aligned peptides is shown. It is clearly observed that the alignment is not optimal even though the core motif (PxxP) is contained by all the peptides. Some of the peptides aligned correctly (class I and II motifs are colored as blue) but others aligned incorrectly (class I and II motifs are colored as red). This suboptimal alignment is produced due to proline-rich property of the peptides and hence predictive performance of any computational model that uses an error-prone initial alignment will severely be affected.

Identifier (MUSI), to address multiple binding specificities of a modular domain [192]. In this method, all the peptide sequences for training purpose are first aligned without any internal gaps and a mixture model is then built to identify multiple PWMs from the aligned sequences [188]. The parameters for the mixture model are optimized by standard Maximum Likelihood with the Expectation-Maximization (EM) algorithm [188, 193]. The number of PWMs (K) is automatically determined by the algorithm.

However, all the PWM-based methods inherently require an initial multiple sequence alignment of the bound peptides, which is a very hard task for poly amino acid sequences. Even minor alignment errors typically cause significant noise in the PWMs and eventually produce error-prone models. Moreover, these complex models for predicting domain-peptide interaction sometimes provide over-specific results, which do not reflect relevant biological insights [187, 194].

### Alignment issues: a review on proline-rich peptide alignment problem

Since PWM-based methods rely on an error-prone peptide alignment process (especially aligning proline-rich peptides bound by the SH3 domains), one risks to introduce a significant amount of noise and therefore obtain under-performing prediction models. SH3 domains typically recognize proline-rich regions of a binding protein. Alignment of the proline-rich peptides is a difficult task (see details in Figure 2.5). Unfortunately, no alignment algorithms are available that can successfully handle this problem. Hence, computational methods based on pre-alignment of proline-rich sequences produce suboptimal models. For example, MUSI tool requires an initial alignment of the binding peptides to build multiple PWM-based models [192]. To see how it affects the predictive performance, we performed an experiment described below.

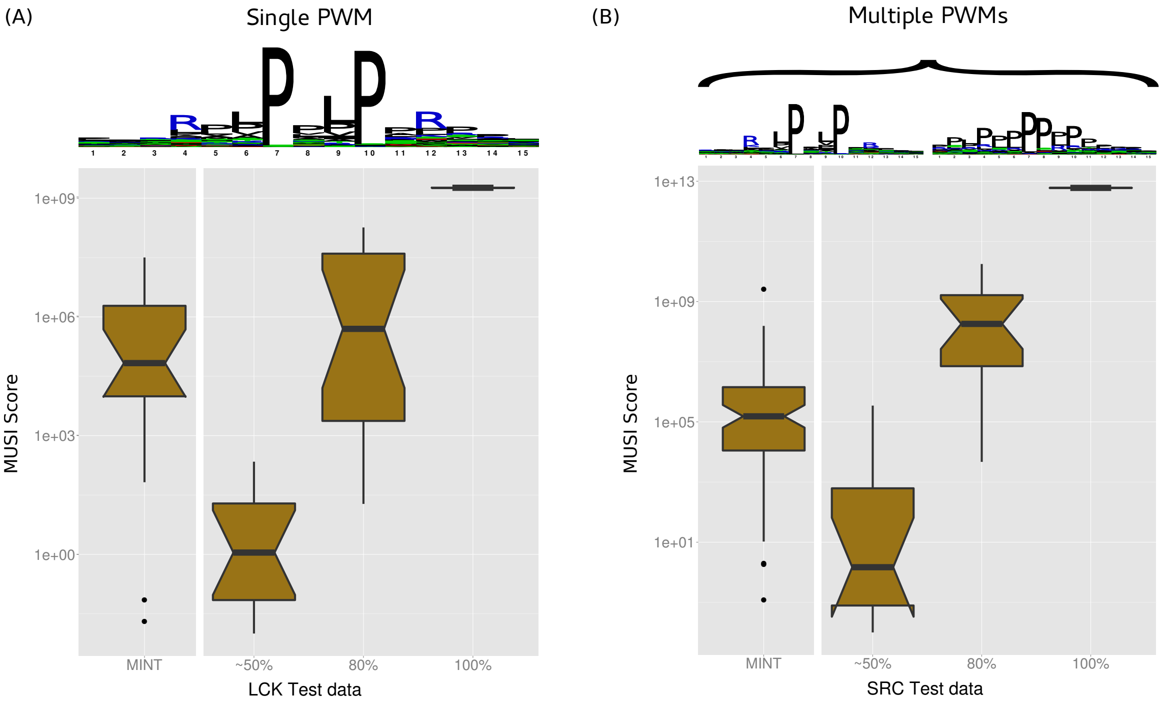


**Figure 2.6.:** (A, B) Whisker-plot of `MUSI` score with different test data for two different human SH3 domains (LCK and SRC). It uses four different test sets for each SH3 domain. In both cases, the first dataset comprises all the known interactions retrieved from the `MINT` database and other three datasets comprise random peptide sequences that are ~50%, 80%, and 100% proline-rich, respectively. `MUSI` produces single PWM for LCK domain and multiple PWMs for SRC domain. It is clearly observed that `MUSI` scores are affected by the proline-rich sequences and thus 100% proline-rich peptides, which are probably all non-binding peptides, achieve highest `MUSI` scores.

`MUSI` models have been trained for human SH3 domains from LCK and SRC protein. The interactions data was derived from a high density peptide array experiment [58]. LCK training data contained 49.33% class I and 50.47% class II peptides, whereas SRC training data contained 52.67% class I and 46.68% class II peptides (see Table A.2.1). Thus, observing multiple specificity (class I and class II) for both domains was expected by the `MUSI` algorithm. But with default settings, `MUSI` produces a single PWM for LCK and multiple PWMs for SRC. Unfortunately, neither of the PWMs define the correct multiple specificity of these domains. In the Figure 2.6, the sequence logos indicate the specificity identified by `MUSI`. To evaluate the predictive performance, we created four different test sets. The first test set was taken from the `MINT` database, which is a high-quality manually curated database [196]. Other three artificial test sets comprise random peptides with ~ 50%, 80%, and 100% proline-rich sequence, respectively. Surprisingly, the peptides achieve higher `MUSI` scores as the percentage of proline-rich increases. In both cases, 100% proline-rich peptides, that are probably all negative data, achieve highest `MUSI` score (see Figure 2.6). This result let us conclude that the PWM-based approaches that depend on pre-alignment of binding peptides always produce suboptimal models, specifically when the binding peptides are rich in one type of amino acid.

### 2.3.2. Machine learning-based methods

Traditional PWM-based methods rely on generative approaches. More specifically, the probability of a PWM-based model is estimated only on positive interaction data where the information on the negative interaction data (non-interacting peptides) is completely ignored. Machine learning algorithms that rely on both positive and negative data to generate discriminative models have an advantage over generative ones. The ability to use positive and negative data allows a discriminative model to identify the decision boundary for the relevant regions of the data space [197]. Previous research showed that the models generated by good quality interacting (positive) and non-interacting (negative) data achieve better prediction quality [198, 199]. Machine learning is a computational method that is widely used for domain-peptide interaction prediction. Mainly supervised machine learning is used to predict modular domain mediated interactions, though there are other forms of machine learning approaches available to deal with different problems (e.g., clustering, affinity prediction etc.). Supervised machine learning methods are separated into two major steps: (i) training and (ii) testing. In the training phase, the data is systematically encoded as a set of feature vectors. This encoding then allows machine learning algorithms (e.g., SVM, Bayesian network etc.) to model relations between features and their respective classes. A validation set is used to optimize the parameters. The goal of machine learning algorithms is to distinguish the positive class from the negative class, and finally apply this knowledge to classify previously unseen data in the test phase. Although there are many strategies that have been used for estimating the performance of a predictive model, cross-validation technique is more commonly used one (see Section 3.6.1 for details).

In the following sections, all the important tools based on machine learning that have been applied to predict modular domain mediated interactions are discussed.

#### SVM-based methods

Support vector machine (SVM) is a well-established computational learning method that recognizes patterns from the training data and builds a discriminant function that separates binding and non-binding interactions. Over the years, several methods based on SVM have been used for prediction of domain-peptide interactions. In 2011, Bader and his group published a semi-support vector regression (SemiSVR) based framework using quantitative positive and qualitative negative training data to predict PDZ-peptide binding affinity [198]. Hui *et al.* employed support vector machines (SVMs) to predict PDZ-peptide interactions from multiple organisms [199]. In their study, they tried to improve the quality of negative data. Recently, an SVM-based method called `DomPep` has been proposed by Li *et al.* for predicting the binding partners of SH2 and PDZ domains. This method applied a nearest neighbor approach (based on domain sequence identity and ligand binding specificity) to extend the training set for each domain, achieving higher domain coverage [200]. Other PDZ-peptide interaction prediction methods based on SVMs are also reported [201, 202]. SVM-based predictors that incorporate discriminative features not only from sequence information but also from structural information have also been proposed by several groups [203, 204]. Hui *et al.* developed an SVM-based method that incorporates the structure-based features that have important roles in protein structure stability and facilitating protein-protein interactions [203]. Hawkins *et at.* proposed threading techniques that generate structure-based sequence alignment for contact residue inference, and used a geometric method to encode the structure of the binding site [204].

### ANN-based methods

Artificial neural network (ANN), which is inspired by animal nervous systems, is a machine learning approach that commonly used to predict numeric quantities, and has been successfully applied to solve pattern recognition problem.

For predicting domain-peptide interactions, several ANN-based methods have been reported. Miller *et al.* developed an ANN-based model to build an atlas of consensus sequence motif of phosphotyrosine dependent binding domains, i.e., SH2 domain [191]. Ferraro *et al.* proposed a machine learning method based on neural network to predict SH3 mediated protein interactions. The features were encoded using the information from known domain-peptide complex structures [205, 206]. Recently, a combined method with ANN and SVM based on sequence and structural information has been proposed to build the model for PDZ domain mediated interactions. ANN was used to make a model that predicts the probability distribution of the contact residues and subsequently this model was used by SVM for predicting binding and non-binding interactions [204]. In this study, the authors showed that the prediction of PDZ domain-peptide interactions can be improved specifically for low sequence similarity domains. Additionally, their method also achieved a better false positive rate. One disadvantage of ANN is that the produced structure is opaque and cannot provide any meaningful information to understand the nature of the solution, although there are some techniques that can produce understandable insights from the structure of neural networks [207].

### Bayesian model-based methods

Bayesian network is a well established machine learning method that can be used to solve variety of computational problems. There are two components, which define the construction of a learning algorithm for Bayesian networks: (i) a function that evaluates the given network based on the data and (ii) a method that searches through the space of possible networks [207].

Without any exception, this method has also been used for prediction of modular domain-peptide interactions. Chen *et al.* proposed a Bayesian model that incorporated structural information of a reference PDZ-peptide complex structure (PDB id: 2PDZ) to predict PDZ-peptide interactions [88]. They identified 38 position pairs involving 16 positions in the

PDZ domain and 5 positions in the peptide. All the position pairs were then incorporated as features.

Other machine learning models to address modular domain-peptide interactions have also been reported previously [208, 209]. However, all the machine learning-based methods proposed to date have some severe caveats that affect their performances. We will discuss several of these problems in the following section.

### Modeling issues: a review on the imbalanced dataset problem

From an *in silico* modeling point of view, a key characteristic of the problem at hand is that the available supervised information on peptide binding induces imbalanced datasets, i.e., for certain SH2 domains, information on real interactions can be up to 15 times more abundant than information on the lack of interaction (see Table A.1.1). In literature, it is known (see [210] for a recent survey) that severe imbalanced class distributions negatively affect the performance of machine learning approaches. The exponential increase in the number of publications dedicated to imbalanced data management in the last decade is a clear indication of the importance of the issue.

The problem arises since mainstream machine learning algorithms are not designed to compensate for skewed class distributions, and concentrate on being accurate only on the majority class. Two major causes of problems with class imbalance are: (i) the choice of an adequate performance measure to guide the selection of the best hypothesis and (ii) the discrepancy in the data distribution between the model induction (train) and the model application (test) phase [211].

To illustrate point (i), consider a typical protein interaction prediction problem: while the number of possible interactions grows quadratically with the number of proteins, the number of positive interactions grows typically only linearly (i.e., one protein will bind to a small fixed number of other proteins). In this case, the standard accuracy measure is not appropriate, since a rational choice based on maximizing the predicted accuracy (in an equal cost scenario) would inevitably be biased towards the majority case, and hence the algorithm will almost always predict a negative/no-interaction response. To deal with this issue, many techniques have been developed that try to explicitly and differently model the cost of each type of mistake. A major drawback of this approach is that the optimal cost matrix is unknown and the result is therefore highly dependent on expert knowledge and a set of arbitrary/heuristic choices.

As for point (ii), it has been recognized that the issue is linked to the *within-class imbalance* problem and the *small disjuncts* problem [212]. The phenomenon arises when the class concept is composed of many sub-concepts/sub-clusters, each represented by relatively few examples. Standard approaches achieve suboptimal results here, since not enough examples are available to model an adequate response for these exceptional although significant cases.

Standard approaches are further compromised, if the sampling procedure in the test phase differs from the one used to collect the training set. This typically happens when a small sub-cluster in the training set is over-represented in the test set (e.g., if cellular conditions or experimental parameters change during data collection).

However, some guidelines are emerging in the machine learning literature on how to counter-balance the *small-disjuncts* problem; the main recommendation is to prefer intelligent over-sampling techniques to down-sampling as the latter always implies a loss of information, which ultimately results in under-performing models. General approaches to over-sampling, such as the popular synthetic minority oversampling technique (`SMOTE`) [213], have the drawback of requiring an explicit instance representation (generally in some vector space of relatively low dimensionality), and are therefore more difficult to adapt to the type of data typically encountered in bioinformatics applications (i.e., sequences or graphs).

### Learning issues: a review on the semi-supervised problem

The task of estimating when an existing peptide belongs to the non-interaction class can be viewed as a special instance of the well-studied semi-supervised learning task (SSL) [214], i.e., learning from a small amount of labeled data and a large amount of unlabeled data. Here, differently from the general problem formulation, the main idea is to use the unsupervised material to have a better characterization only of the minority class; in our case, the one representing the absence of protein-peptide interaction.

Several strategies have been developed to deal with the SSL problem, such as self-training, expectation maximization (EM) with generative mixture models, co-training, transductive support vector machines, and graph-based methods. In order for SSL methods to use effectively the small amount of labeled data, strong model assumptions need to be made. Note that this is a critical step, as it has been observed that if the model assumptions are not matching the nature of the problem, then using unsupervised material hurts the predictive performance [215]. Therefore, one should review the assumptions made by each SSL strategy, matching them to the specific application case.

Expectation maximization techniques with generative mixture models can be used when data is well clustered according to the class information. In our case, clustering peptides using a metric that makes use of all amino acid information does not induce a good class separation, in fact it is believed that binding is the result of the joint presence of only a few specific amino acids in specific positions.

Co-training is used when features naturally split into two sets, with a different instance coverage, but this is not the case for our application.

Graph-based methods perform a type of information spreading on unsupervised instances that is meaningful when two nearby instances (i.e., instances with similar features) tend to be in the same class. For the same reasons detailed for the EM case, this type of bias is not appropriate for our application.

Finally, the self-training approach, which relies only on the good discriminative properties of the base classifier. The method is a simple wrapper scheme around a base classifier: the initial labeled data is used to train the classifier which then assigns a label to the remaining material. This training procedure may continue for several iterations depending on the dataset. In our application, this is the most suited approach that could successfully handle the semi-supervised problem, specifically for SH2 and PDZ domains (see Section 3.2.3 and 3.4.4 for details).

#### Reliable negative data issue: a review on high-confidence negative data problem

Reliable training data is essential to build a good discriminative model. While positive training data can be obtained from different experiments and/or literature, unfortunately enough reliable negative data is often unavailable. Datasets derived from the high-throughput experiments usually suffer from the same problem. For example, phage display experiment only provides positive interaction data. Thus, generating artificial negative data to build powerful computational models is an open challenge. Previous research showed that proper selection of artificial negative data increases the performance accuracy of a predictor [216, 217]. In common practice, random and shuffled peptide sequences have been used for generating artificial negative instances. However, in the prediction of protein-protein interaction, the randomly shuffled peptide instances produce models with lower prediction accuracy since they do not resemble real biological sequences [217], and are not useful for determining meaningful class boundaries. Lack of high quality and biologically relevant negative (i.e., non-interacting) data is therefore one of the biggest drawbacks of most of the available machine learning models that predict modular domain-peptide interactions.

#### Overfitting issues: a review on the regularization problem

Instead of capturing underlying trend of the data, overfitting occurs in a machine learning algorithm, when it captures the noise in the data. More specifically, if a model fits the data too well, it results in overfitting and produces a poor predictive model. To improve the predictive performance, an appropriate technique (e.g., regularization, pruning etc.) should be used that can counter balance the over-training phenomena, i.e., the tendency to specialize the model on the specific training data idiosyncrasies. A regularized predictor is more robust to noise and offers guarantees of a better predictive behavior on unseen instances. It is an unfortunate state of affairs that this aspect is often ignored in the development of novel bioinformatics systems.

### 2.3.3. Structure and energy-based methods

While above approaches rely only on sequence information of domain and peptides, several other approaches have also been reported that exploit binding information of domain-peptide complex structures. Structural information of domain-peptide complexes are very important to understand the binding specificity of respective modular domains and thus using these features can increase the predictive performance of a method. Moreover, they are often capable of distinguishing between the residues that prevent binding and the residues that are not favored at the binding site, which strengthen the prediction quality.

Over the past several years, plenty of approaches that derive energy models using structural information of domain-peptide complexes to address the specificity of modular domain mediated interactions have been reported. The 3D-QSAR based comparative molecular field analysis (CoMFA) was developed by Lee *et al.* to investigate the quantitative structural activity relationship for SH2-binding peptides [218]. Sánchez *et al.* proposed a method to predict SH2 domain-peptide interactions using SH2-phosphopeptide complex structures and FlodX algorithm [219]. In this method, structure-based energy function was used to calculate the energy of a protein complex. Protein backbone sampling was used by Smith *et al.* to predict the sequence space of peptides that are recognized by PDZ domains [220]. Kaufmann *et al.* proposed an optimized energy function to predict the binding specificity of 12 PDZ domains [221]. More recently, Hou *et al.* proposed a structure-based method that uses molecular interaction energy components (MIECs) for characterizing residue-residue interaction pattern between SH3 domain and interacting peptides [222]. One structure-based method to identify the specificity of SH3 domains using *in silico* mutagenesis has been reported by Fernandez-Ballester *et al.* [223]. Other structure-based methods have also been previously reported to address the specificity of modular domains [224–229]. Unfortunately, these structure-based approaches essentially rely on solved 3D domain-peptide complex structure, which are, however, known only for a few cases, and are also computationally very expensive. Moreover, most of the structure-based methods typically cannot make use of the available high-throughput data.

One exceptional work proposed by Wunderlich *et al.* in 2009, who studied the physical origin of SH2 domain-peptide specificity by integrating structural information with a quantitative high-throughput domain-peptide interaction dataset and developed an energy model to accurately predict SH2 domain-peptide interactions [230]. Three different methods were described to construct an interaction map: (i) information-based, (ii) structure-based, and (iii) hybrid-based method. They found the amino acid positions in the peptides and domains that confer specificity of the interactions by using information-based and structure-based methods. This method can also be applied to the SH2 domains or pY peptides that are not used in model construction. However, the good performance reported seems to be due to some over-training issues (see Section 4.2.6 and Figure A.1.4).

### Domain coverage issue: a review on lack of domain coverage problem

While the experimental data has become invaluable to build powerful computational models to describe the specificity landscapes of modular protein domains, this data may be rich only for certain domains while it is scarce or completely missing for others. Most of the methods that have been developed till date use domain specific or single domain models, meaning developing models for those domains that have less or no experimental evidence is nearly impossible. It is already known that the different domains in the same specificity family share their conserved binding properties. Hence, instead of building single domain models, building multiple domain or family-based models would be a more attractive strategy. Domain coverage of a model can be increased by combining the experimental data for domains with similar binding preferences, which includes orthologous domains from other organisms as well.

Such methods have recently been reported for a few modular domain family. For example, Chen *et al.* developed a Bayesian model that can predict the interactions for any PDZ domains [88]. In this model, the PDZ domain-peptide structural information have been incorporated into the features. Bader and his colleagues have also developed SVM-based approaches to solve the same problem [199, 203]. A model for all SH2 domains is also available [230]. But these methods rely on a single domain-peptide complex structure, which is oversimplification of the diverse specificity problems of modular domains and thus do not perform well for all domains. Recently, Li *et al.* published a sequence-based method where they combined the domain information using nearest neighborhood approach to extend the training set for each domain. Finally, they built models that cover 174 PDZ domains and 97 SH2 domains [200].

## 2.4. Discussion

In this chapter, we have discussed several high-throughput techniques that are extensively used for describing the specificity of modular protein domains, and the computational models that have been developed for predicting the modular domain mediated interactions. We have also described the challenges of some important computational methods and their limitations. There are many computational methods available for the prediction of domain-peptide interactions but they are mainly affected by the aforementioned problems. Our methods account for many advantages, as we tried to overcome the limitations of existing computational methods mentioned in the previous section. In Chapter 3, we will describe the strategies that have been taken to circumvent these limitations. The key advantages of our methods are described in Section 3.5.

# Chapter 3

# New approaches for predicting modular domain-peptide interactions

## 3.1. Overview

It is known that the modular domains are highly specific towards their binding ligands. Due to a high number of modular domains, one has to resort to high-throughput data to define the binding specificity of modular domains. Thus, *in silico* ligand binding prediction of modular domain mediated interaction is of great interest. Currently, several computational methods have been proposed to predict modular domain-peptide interactions. However, they have many shortcomings as described in Chapter 2.

In this chapter, we present three efficient and accurate methods for prediction of SH2, SH3, and PDZ domain-peptide interactions using machine learning approaches. All methods are based on support vector machine with different kernel functions ranging from polynomial, to Gaussian, to sophisticated graph kernels. Our methods are successfully capable of dealing with the most challenging problems in this area of research and in contrast to other existing methods, our models have several advantages. This chapter is split into five different sections; first three sections describe the modeling strategies of all three methods, and last two sections describe the key advantages of our models and performance measure techniques, respectively. The work presented in this chapter is a part of the following publications: [P2], [P3], and  [P4].

## 3.2. SH2 modeling

In this section, we present an efficient machine learning method based on polynomial kernel for prediction of SH2 domain mediated interactions. A polynomial kernel allows us to exploit the dependencies between the amino acid positions in the peptide sequences. Additionally, we used a semi-supervised learning approach to deal with highly imbalanced training data. Finally, we present the prediction models for 51 human SH2 domains.

### 3.2.1. Feature encoding

Previous studies showed that residues in the close vicinity of the phosphotyrosine (pTyr) are highly predictive for SH2-peptide binding [181, 183, 230]. For example, it is known that the SH2 domain of CRK binds peptides where amino acid Leu or Pro is in position +3, however, the presence of other amino acids (i.e., His, Arg, Ala, Pro) in position +1 and +2 can inhibit the interaction [43].

Here, we follow the literature and restrict the peptide sequence to 6 specific positions, namely we extract the amino acids in positions ranging from 2 upstream to 4 downstream of the phosphotyrosine residue. A peptide is therefore mapped into a binary vector $x$ living in a $20 \times 6 = 120$ dimensional (the central amino acid is always a phosphotyrosine and is therefore not included in the encoding), that is, for each position, we reserve 20 dimensions (one for each amino acid) and encode the amino acid type with a 1 in the corresponding dimension and 0 elsewhere.

For each domain $D_j$, we compile a dataset encoded as a set of pairs $(x_1,c_1)$,..,$(x_n,c_n)$ where, $x_i$ is the binary feature vector for peptide $P_i$ with the class label $c_i \in \{-1, 1\}$. The class label is +1 if the domain $D_j$ interacts with peptide $P_i$ and -1 otherwise.

### 3.2.2. Predictive model

As a predictive model, we employed a regularized polynomial kernel support vector machine (SVM) [231]. We used the SVM implementation in `C` language provided in `SVM`$^{light}$ [232].

### Polynomial kernel

A polynomial kernel is a kernel function that computes the similarity between training samples (vectors) in the feature space over polynomials of the variables to learn a non-linear model. A polynomial kernel function is represented as $(\langle x, x' \rangle)^d$, which computes the dot product of two vectors: $x$ and $x'$, and raises the result to the power $d$. The polynomial kernel function for degree $d$ is defined in [233] as:

$$K(x, x') = (1 + \langle x, x' \rangle)^d, \tag{3.1}$$

where "1" is a constant, which is needed to consider the effects of all degrees that are less than $d$. Choosing the value for $d$ is important for regularization. $d = 1$, which is a linear model would be a good starting value and can be incremented until the estimated error ceases to improve [207]. Note that for building non-linear model, $d > 1$ is required as $d = 1$ only provides linear model. Polynomial kernel of degree 2 ($d = 2$), which is a quadratic polynomial function, and a feature space with two inputs $x_1$ and $x_2$ can then be defined as:

$$\begin{aligned} K(x, x') &= (1 + \langle x, x' \rangle)^2 \\ &= (1 + x_1 x'_1 + x_2 x'_2)^2 \\ &= 1 + 2x_1 x'_1 + 2x_2 x'_2 + (x_1 x'_1)^2 + (x_2 x'_2)^2 + 2x_1 x'_1 x_2 x'_2. \end{aligned} \tag{3.2}$$

One of the main hyper-parameters in SVM is the cost parameter or $C$, which provides some flexibility in an enlarged feature space for data separation. Basically, it creates a soft margin that allows you to penalize misclassification of training instances. A large $C$ value will create a irregular boundary or a small margin that can lead to an overfitting situation in the original feature space. Conversely, a small $C$ value will create a large margin even if that hyperplane misclassifies more points.

### Regularized non-linear support vector machine

Predictive systems based on PSSMs are essentially linear classifiers. To see why, we review the design principles for the state-of-the-art PSSM system `SMALI` [183].

**PSSM-based `SMALI` model** In `SMALI`, a procedure is employed to compute a weight matrix $S_{r,c}$ with $r = 6$ rows and $c = 19$ columns (the Cys amino acid is not represented). The domain-peptide interaction is predicted computing a score value as:

$$s(x) = w^T x, \tag{3.3}$$

where $x$ is a 114 dimensional vector constructed as specified in Section 3.2.1, $w = vec(S^T)$ where $vec$ is an operator that transforms a matrix $M_{r,c}$ into a column vector $v$ of size $r \cdot c$, by concatenating all columns. Peptides scoring above a predefined threshold are classified as binding. In `SMALI`, a relative score is defined in such a way as to have a unit threshold. The relative score is then the ratio between the original score and a reference score $b$. The classifier becomes $s(x)/b \geq 1$, which can be rewritten in a canonical linear form as:

$$w^T x - b \geq 0. \tag{3.4}$$

From a machine learning perspective, the procedure employed in `SMALI` to compute $S$ and $b$ is rather involved and heuristically motivated. The elements in the matrix $S$ are computed from OPAL [183] experimental results, and essentially correspond to the difference between the average position specific counts of specific amino acids in the positive examples minus the overall average counts.[1] These quantities are then transformed so to extract information theoretic quantities as a proxy of the importance (the weight) of each position specific amino acid.

The domain specific reference score value $b$ is defined as the value corresponding to the top $q = 4.5\%$ raw SMALI scores over all human proteins in the UniProt database that contain tyrosine (Tyr). The choice of the fixed value 4.5% was based on two experiments over the domains BRDG1 SH2 and GRB2 SH2, arbitrarily chosen as representative cases. The optimal (w.r.t. F-measure) threshold for the raw SMALI score was computed using a selection of 1488 peptides for BRDG1 (yielding a SMALI value of 1.4) and 720 peptides for GRB2 (yielding a SMALI value of 1.65). The percentiles corresponding to these thresholds were 3.5% for BRDG1 and 5.5% for GRB2. The final value $q = 4.5\%$ was chosen as their average. As a result of all these choices, it is hard to identify a clear objective for which the proposed linear solution should be optimal.

Here, we propose two ways to improve PSSM linear models: (i) upgrading the system from linear to non-linear and (ii) making the system more robust using regularization techniques.

**Regularized non-linear model** Non linear models allow to express decision rules that can differentiate between the joint status of two or more position specific amino acids and the status of the same elements taken independently. In this way, non additive effects can be modeled, for example, consider a case whereby the presence of amino acid Asn in position +2 alone is not sufficient to guarantee the interaction and neither is the presence of amino acid Lys in position -1. However, if these amino acids are occurring in their respective positions at the same time, then the binding occurs. Another type of non-linear effect could raise when the presence of a either one or the other amino acid is sufficient for binding but when they are both present then they interfere with each other and no binding takes place.

As a non linear model, we choose to upgrade the standard linear SVM via a polynomial kernel of the type described in Equation 3.1. To see how a kernel allows an otherwise linear model to become sensitive to multiple interacting amino acids, we briefly review the ideas behind the *"kernel trick"*. Given a linear predictive model $f(x) = \text{sgn}(\langle w, x\rangle + b)$, where $\langle \cdot, \cdot \rangle$ represent the dot product operation, one can employ the support vector machine [231] algorithm to determine the support elements and rewrite the decision function as:

$$f(x) = \text{sgn}(\sum_i y_i \alpha_i \langle x, x_i \rangle + b), \tag{3.5}$$

where the non zero $\alpha_i$ select which, among all $x_i$, are the support vectors. The trick consists now in replacing the standard dot product with a "kernel function" $K(x, x') = \langle x, x' \rangle$, i.e., a function which is symmetric and positive semi-definite [234]. Choosing an appropriate kernel function allows us to transform a linear classifier into a non linear one. Exploiting re-

[1] This corresponds geometrically to find the difference vector between the center of mass of the positive set and of the overall set. Had it been the difference vector between the center of mass of the positive set and of the negative set, it would have resembled the well known Fisher discriminant model.

sults known from the Reproducing Kernel Hilbert Spaces theory, one can equate the choice of a kernel function to the selection of an appropriate feature mapping function $\phi : X \mapsto \mathbb{R}^d$ and write $K(x, x') = \langle \phi(x), \phi(x') \rangle$. It is often possible to efficiently compute $K(\cdot, \cdot)$ without having to compute $\phi(x)$, i.e., without having to represent the instances explicitly in the transformed feature space. This is particularly beneficial when the size of representation is very large (it can also be infinite in the case of Gaussian kernels). One such case is the polynomial kernel; to fix the ideas, we provide the explicit mapping of a quadratic kernel $K(x, x') = \langle x, x' \rangle^2$, which in the simple case of two dimensional instances would result in $\phi : \mathbb{R}^2 \mapsto \mathbb{R}^3$, e.g., $(x_1, x_2) \mapsto (x_1^2, x_2^2, x_1 x_2)$.

In our domain, this means that with a quadratic kernel we can model interactions between any of two positions in the peptide. Note that in the general case, one can account for all interactions of order $d$ by employing a polynomial kernel of degree $d$, without having to explicitly enumerate all combinations. In our case, with $N = 120$ and a polynomial of degree $d = 3$, we are implicitly working in a vector space with 300K dimensions. Here, the number of different monomials of degree $d$ for $N-$dimensional vectors can be computed as:

$$\binom{d + N - 1}{d} = \frac{(d + N - 1)!}{d!(N - 1)!}. \tag{3.6}$$

To further improve the predictive performance, we propose to use regularization techniques to tackle the over-training phenomena, which is often ignored in the development of novel bioinformatics systems. In practice, a regularized predictor is more robust to noise and offers guarantees of a better predictive behavior on unseen instances. Among the several ways to ensure a regularized solution, we adopt the strategy championed in SVM, i.e., we minimize the complexity of the model by constraining the size of $w$ and the degree of the polynomial $d$. We do this using a cross-validation procedure in order to achieve a good compromise with respect to the training misclassification error. In practice, the SVM optimal hyperplane is determined as the solution to a minimization problem where the objective function combines a term proportional to the training error and a term proportional to the complexity of the model (computed as the norm of the hyperplane coefficient vector). The mixing coefficient that weights the importance of the error w.r.t. the model complexity and the degree of the polynomial kernel are selected from a finite set of alternatives. The best parameter combination is chosen by evaluating the predictive performance of each specific model over a held out set of instances (the validation set). Note that the performance of the selected model is evaluated over a further held out set of instances (the test set) that has never been used neither in the training phase nor in the validation phase.

### 3.2.3. Negative class definition

Using the polynomial kernel, we achieve a higher SH2-peptide interaction modeling flexibility. As a consequence of this increased flexibility, we need a larger number of training instances. Notwithstanding the availability of dataset derived by high-throughput techniques, we still suffer from a lack of reliable negative data. This is the main cause for the high imbalance: for some SH2 domains, information on real interactions can be up to 15 times more abundant than information on the lack of interactions (see Table A.1.1). It is known that in these conditions predictive systems produce suboptimal results [216, 217, 235]. To mitigate these issues, we have employed a semi-supervised learning approach (SSL) [214]. The general strategy of SSL is to learn from a small amount of labeled data and a large amount of unlabeled data. Our proposed pipeline is depicted in Figure 3.1. The main idea is to bootstrap from a smaller set of reliable negative instances and only select peptides that we are highly confident to yield negative interactions. More specifically, the pipeline works as follows: (i) an initial high quality, experimentally verified dataset is extracted from high density peptide arrays and micro array results; (ii) data is rebalanced using a self-training strategy with polynomial SVM; (iii) model selection is performed to select the best model complexity for each specific SH2 domain. The key points here are: (a) rebalancing strategy and (b) self-training phase. For rebalancing, we use over-sampling in order to not throw away valuable information as would be done with under-sampling strategies. In over-sampling method, the instances from the minority class are duplicated randomly until the class balance is adjusted. The self-training approach relies only on the good discriminative properties of the base classifier. The method is a simple wrapper scheme around a base classifier: the initial labeled data is used to train the classifier which then assigns a label to the remaining material. The most confident predictions are then iteratively added to the training set and the classifier is re-trained. The method name derives from the fact that the classifier uses its own predictions to teach itself. The bias is now adequate if the base classifier can learn the importance of each combination of amino acids in specific positions. In our case, the confidence is scored as the distance from the discriminative hyperplane.

### 3.2.4. Model fitting protocol

The model parameters that can be tuned are the polynomial degree $d \in \{1, 2, 3\}$, and the cost parameter $C \in \{0.01, 0.1, 1, 10\}$ used to trade-off generalization for data fitting.

In order to estimate the expected predictive performance for our approach, we computed the 5 measures: sensitivity, specificity, precision, area under the receiver operating characteristics curve, and area under the precision recall curve (see Section 3.6.2 for details) under a *stratified* 5-fold cross-validation scheme.

In particular, all the available data is partitioned into 5 parts ensuring the same proportional distribution of positive and negative instances in each part. Each part is used in turn as a held out validation set, while the remaining 4 parts are used as training set.

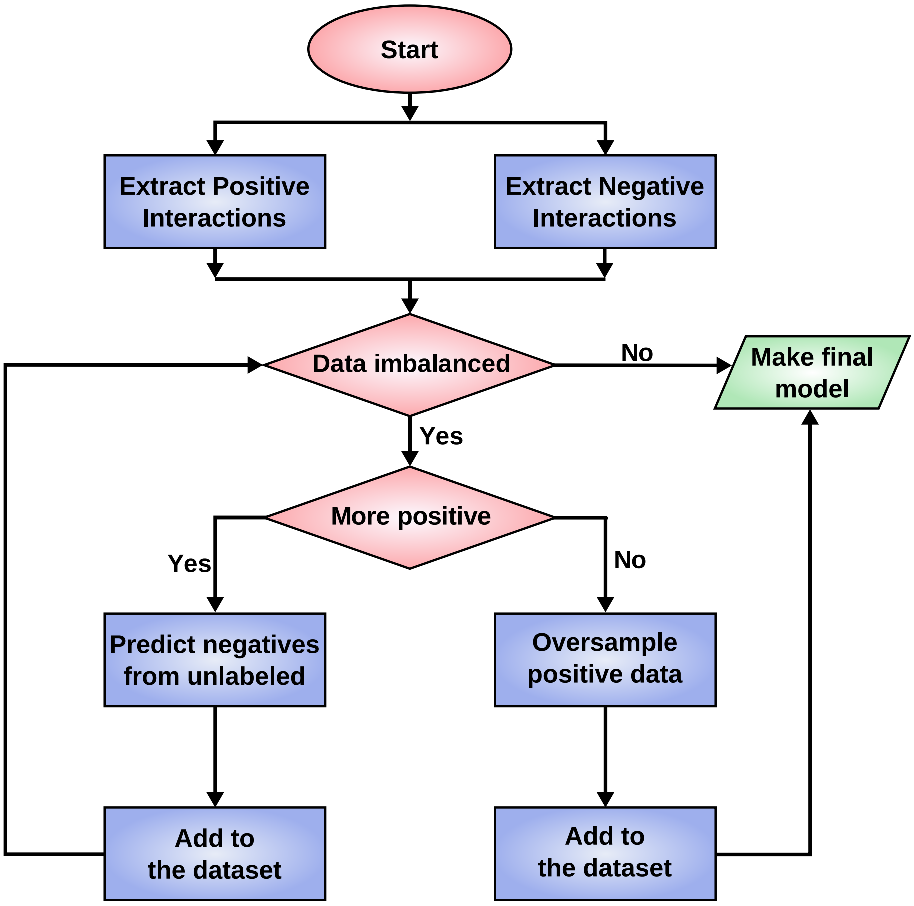


**Figure 3.1.:** Flowchart for the iterative negative data filtering. An initial high quality dataset is extracted from experimental evidence. If the negatives are in excess (right branch), then we simply duplicate the positive instances. If the positives are in excess (left branch), then we make an initial model using over-sampled negatives; this model is then used to score all the available peptides. Those that are more confidently predicted as negatives are added to the dataset. The procedure is iterated until a balanced dataset is reached. The final model is computed on the balanced dataset. The figure is taken from [P4].

We determined the optimal parameter configuration (i.e., the pair $(d, C)$) as the minimum of a 10-fold cross-validated AUC ROC measure for each of the 5 training sets, independently. We then selected the most frequent parameter configuration pair $(d, C)$. This was the configuration finally used in the *stratified* 5 fold cross-validation.

We also performed 10 repetitions of a 75% - 25% random split of the available data to create 10 train/test datasets. We proceeded in an analogous fashion (10-fold cross-validation) to determine the most frequent parameter configuration pair $(d, C)$. The final average performance estimate is comparable to that obtained in the 5-fold cross-validation setting (see Figure A.1.1).

## 3.3. SH3 modeling

In this section, we present an effective machine learning method for the prediction of SH3 domain-peptide interactions. The method is based on a graph kernel approach that, in contrast to the majority of other approaches, does not require the peptide sequences to be aligned and can, at the same time, exploit higher-order dependencies between amino acid residues. Furthermore, a generative approach is used for false negative refinements. Finally we show how to build a model that takes as input both the peptide information and the (aligned) domain amino-acid sequence. By doing so, we can exploit information from related SH3 domains and enhance the overall prediction performance.

### 3.3.1. Feature encoding

For some protein domains, it is possible to identify a key amino acid necessary for a successful binding of a peptide (e.g., the phosphotyrosine for the SH2 domain). This pivotal amino acid can then be used to identify an absolute reference system that allows to represent the peptide as a fixed size vector, i.e., each amino acid is identified as having position $+i$ or $-i$ starting from the pivotal amino acid. For SH3 domains, the situation is, however, more complex as the key amino acid (proline) is abundant throughout the peptide sequence. A unique reference system based on proline cannot, therefore, be easily identified. Commonly, an initial alignment of the peptide sequences is performed in a pre-processing step. Errors in this phase can lead to a bad estimate of the model's parameters and ultimately to bad predictive performances. To circumvent this issue, we employed a sophisticated kernel approach, which eliminates the need of an initial peptide alignment for building predictive models. We built two different predictive models: (i) single domain model and (ii) multiple domain model.

#### Single domain feature encoding

Here, we propose a kernel approach defined independently of an absolute reference for amino acid positions. In this way, we can move from a fixed-size vector type of encoding to a variable length sequence type encoding while still preserving a high discriminative power. The shift from a vector based to a sequence based approach can be extended further: if we move from sequences to graphs, we can then encode any other ancillary information on specific amino acids. To do so, we have to move from string kernels to efficient graph kernels. To ensure low run-times, we resort to the recently introduced [236] Neighborhood Subgraph Pairwise Distance Kernel (`NSPDK`) (see Section 3.3.2).

In more detail, in order to encode the peptide information, we proceed as follows. Given the experimental CHIP design constraints in peptide array library, we can only use peptide sequences of exactly 15 residues in length. We enrich the information available on each amino acid with their average physico-chemical properties, i.e., charge and hydrophobicity.

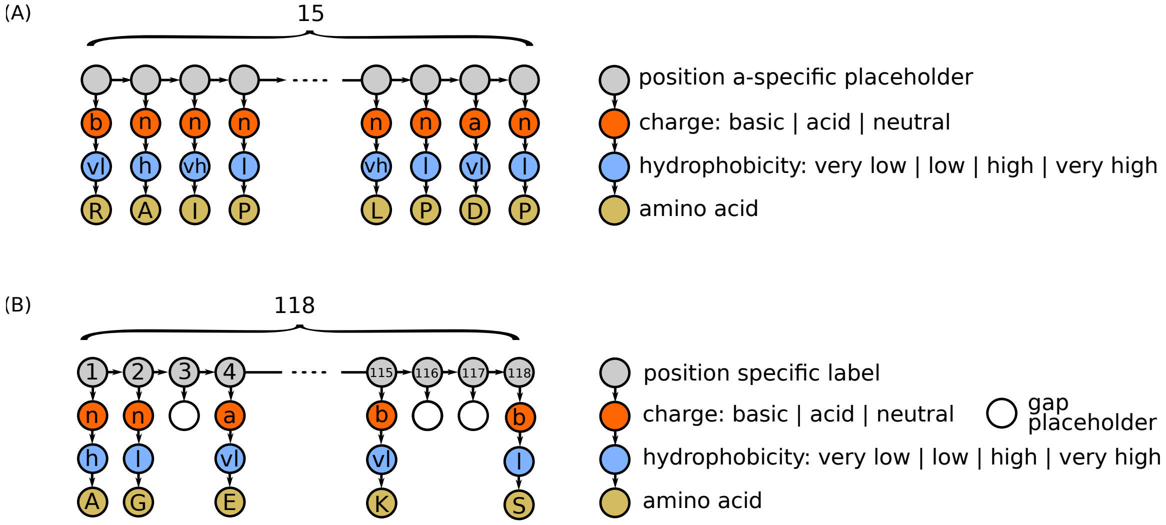

(A)
15
b
n
n
n
n
n
a
n
vl
h
vh
l
vh
l
vl
l
R
A
I
P
L
P
D
P
position a-specific placeholder
charge: basic | acid | neutral
hydrophobicity: very low | low | high | very high
amino acid
(B)
118
1
2
3
4
115
116
117
118
n
n
a
b
b
h
l
vl
vl
l
A
G
E
K
S
position specific label
charge: basic | acid | neutral
gap placeholder
hydrophobicity: very low | low | high | very high
amino acid

**Figure 3.2.:** (A) Graph encoding for peptide sequences where amino acid positions do not have an absolute reference, since we have eliminated the need for an error-prone initial peptide alignment. (B) Graph encoding for domain sequences where amino acid positions receive an absolute positional reference according to a consensus domain alignment. Gaps receive a special encoding. In both cases (A and B), the encoding is enriched with charge, hydrophobicity, and amino acid-type information. The figure is taken from [P3].

Since the graph kernel approach can deal only with discrete labels, we discretize all properties. More specifically, as for charge, we have divided all common 20 amino acids into 3 groups as basic (R, K, H), acidic (D, E), and neutral (the remaining amino acids); as for hydrophobicity, we have identified 4 groups (very low, low, high, and very high) based on their hydrophobicity scales following [237], obtaining: I, L, V as very high hydrophobic residues, A, M, C, F as high hydrophobic residues, G, T, S, W, Y, P as low hydrophobic residues and rest of the amino acids (i.e., R, K, H, D, E, Q, N) are considered as very low hydrophobic residues.

The peptide is then modeled as a chain of unlabeled vertices: one per amino acid. Each vertex is then connected with a side chain graph that encodes the ancillary properties, namely, in order of proximity: the charge, the hydrophobicity, and the amino acid code (see Figure 3.2.A). In order to generate features that are discriminative of the sequence direction, we model the peptide as a directed graph.

### Multiple domains feature encoding

When developing models for single domains, the input encodes only the information for the peptide sequence. However, when we want to induce a general model for a subset of related domains, the input should include also information on how a specific domain relates to the other ones, so that useful knowledge can be transferred from interactions on similar domains. To do so, we model the domain amino acid sequence information in a similar fashion to the peptide encoding with one important difference: since the position of specific amino acids is relevant to determine the specificity of the domain-peptide interaction, we additionally encode the information of an absolute positional reference. In order to do so, we align the related domains with the MUSCLE [42] alignment software. Note that in contrast to the peptide alignment, the SH3 domain alignment is highly reliable; mainly the alignment of n-SRC-loop and RT-loop of the domains. Each domain specific sequence is then projected onto the alignment and the necessary gaps are finally introduced (see Figure 3.2.B). In this strategy, we pool all the related domains and their binding peptides into one set to keep the information for the both peptide and domain sequences (see *Multiple Domains Modeling* in Section 3.3.4 for details). The input for the multi-domain model is therefore comprised of two disconnected components, one for the peptide and one for the domain. In order to eliminate ambiguity issues, we distinguish the label alphabet for the peptide sequence from that of the domain sequence by means of appropriate prefixes.

### 3.3.2. Graph kernel approach

In the past decade, machine learning and data mining community have made progress to allow more flexibility in the input data type formats, and extended it from fixed size vector to more flexible and dynamic formats starting from sequences, to trees, and eventually to graphs. In supervised learning method, one advantage is that a linear model with good generalization properties can be easily extended to a non-linear model using the *kernel trick* [238, 239]. Various kernel functions are available (e.g., polynomial, Gaussian etc.) and commonly used with support vector machine to implicitly map the input data into a very high-dimensional feature space expressed by a suitable dot product. Such a kernel function, namely graph kernel, has been proposed to deal with the entities represented as a graph. It uses the dot product functions and computes the similarity measure between the graphs. Although there are several types of graph kernels available, we have resorted to a fast kernel, namely Neighborhood Subgraph Pairwise Distance Kernel (NSPDK), which has been recently introduced by Costa *et al.* [236], because it is suitable for large sets of sparse graphs with discrete vertex and edge labels.

#### Notation and definitions

A graph $G = (V, E)$ consists of two sets $V$ and $E$, where the elements of $V$ are known as *vertices* and the elements of $E$ are known as *edges*. Each edge is associated with a set of two elements of $V$, which are called its *endpoints*. We can denote the *endpoints* by concatenating the vertex variables. For example, $uv$ represents the edge between the two vertices, $u$ and $v$. When $G$ is not the only graph to be considered, the notation $V(G)$ and $E(G)$ are used. A graph is called as labeled graph when the vertex and edge labels are assigned into it, using repetitive symbols from a set of finite alphabet. We denote the function as $\ell$, which maps vertex/edge to the label symbol. Two graphs, $G_1 = (V_1, E_1)$ and $G_2 = (V_2, E_2)$, are isomorphic, if there is a bijection $\phi : V_1 \to V_2$ and can be denoted by $G_1 \cong G_2$. Basically, an isomorphism is a structure-conserving bijection. Thus, two labeled graphs can also

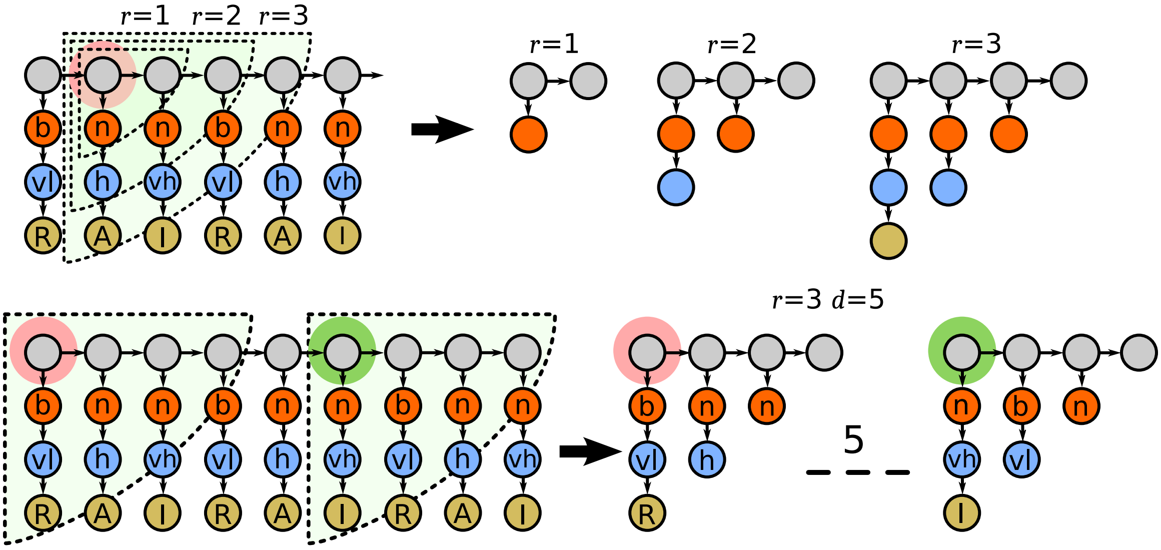


**Figure 3.3.:** Top: NSPDK features for Distance $(d) = 0$ and Radius $(r) = 1, 2, 3$ relative to a given root vertex highlighted in pink. The directedness property of the graph allows to induce features that can differentiate strand directions. Bottom: Example of feature for $r = 3$ and $d = 5$ capable to capture the dependency between two amino acid at relative distance 5. The sequence information that is not contained in the neighborhoods is ignored; the effect is equivalent to a *don't care* pattern. The figure is taken from [P3].

be isomorphic, if there is an isomorphism, which preserves the label information too, i.e., $\ell(\phi(v)) = \ell(v)$. A graph invariant or isomorphic invariant is a graph property, which is identical for two isomorphic graphs.

### Graph kernel

The NSPDK is an example of decomposition kernel [240] by which all the possible "parts", defined by a given relation, are operated. In this case, each part is a pair of special subgraphs, which are known as "neighborhood" subgraphs. Here, the key idea is to generate small neighborhood subgraphs of increasing radii $r < r_{max}$ by decomposing a graph. All pairs of such subgraphs are considered as individual features, if their roots are at a distance $(d)$ not exceeding $d_{max}$ $(d < d_{max})$. We consider the fraction of features in common between two graphs as the similarity notion. The formal kernel definition is reported here. The relation between neighborhood subgraphs is defined as:

$$R_{r,d} = \{(N_r^v(G), N_r^u(G), G) : d(u, v) = d\}, \tag{3.7}$$

where $R_{r,d}$ (neighborhood pair relation) identifies a pair of neighborhood subgraphs of radius $r$, which has root distance exactly equal to $d$. On this relation $R_{r,d}$, one decomposition kernel $\kappa_{r,d}$ is defined as:

$$\kappa_{r,d}(G,G') = \sum_{\substack{A,B \in R^{-1}_{r,d}(G)\\ A',B' \in R^{-1}_{r,d}(G')}} \xi(A \cong A') \cdot \xi(B \cong B'), \tag{3.8}$$

where $R^{-1}_{r,d}(G)$ is the inverse of $R_{r,d}(G)$, which indicates all the possible pairs of neighborhood subgraphs of radius $r$ and the root distance $d$ that exist in the graph $G$, and the indicator function and the isomorphism between graphs are denoted by $\xi$ and $\cong$, respectively. The isomorphism check is performed with the techniques as detailed in *Graph invariant*, later in this section. In our case, amino acid sequences are considered as NSPDK features (see Figure 3.3). The non-normalized NSPDK is defined as:

$$K(G,G') = \sum_{r}\sum_{d} \kappa_{r,d}(G,G'). \tag{3.9}$$

For increasing the efficiency, an upper bound on the radius and distance parameters can be imposed as:

$$K_{r_{max},d_{max}}(G,G') = \sum_{r=0}^{r_{max}}\sum_{d=0}^{d_{max}} \kappa_{r,d}(G,G'). \tag{3.10}$$

Finally, a normalized version of $\kappa_{r,d}$ can be defined, that is:

$$\hat{\kappa}_{r,d}(G,G') = \frac{\kappa_{r,d}(G,G')}{\sqrt{\kappa_{r,d}(G,G)\kappa_{r,d}(G',G')}}. \tag{3.11}$$

This ensures that the graph features induced by all values of radii and distances are equally weighted irrespective of the feature space dimensionality.

### Graph invariant

Unfortunately, for solving the graph isomorphism problem (GIP), it is unknown whether polynomial algorithms exist. However, for special graph classes, polynomial algorithms do exist [241]. Few algorithms that are exponential have also been developed to solve the GIP problem previously [242, 243]. Since the exact isomorphism test is computationally very expensive, Costa *et al.* proposed a solution, similar to [244], where the exact isomorphism test is substituted by introducing an efficient graph invariant computation [236]. Here, the key idea is to produce an identical string from two isomorphic graphs by efficient graph serialization procedure. Then, the string can be mapped into an integer code by an iterative hashing technique. Therefore, the isomorphic test can be easily substituted by an equality test between the integer codes of two graphs. This whole process works in two main steps: (i) construction of a graph invariant encoding $\mathcal{L}^g(G)$ and (ii) using a standard hash function $H(\mathcal{L}^g(G)) \to \mathbb{N}$ to get the desired identifier. Note that in general, the process is affected by potential collisions between two non-isomorphic graphs. This can happen either due to the

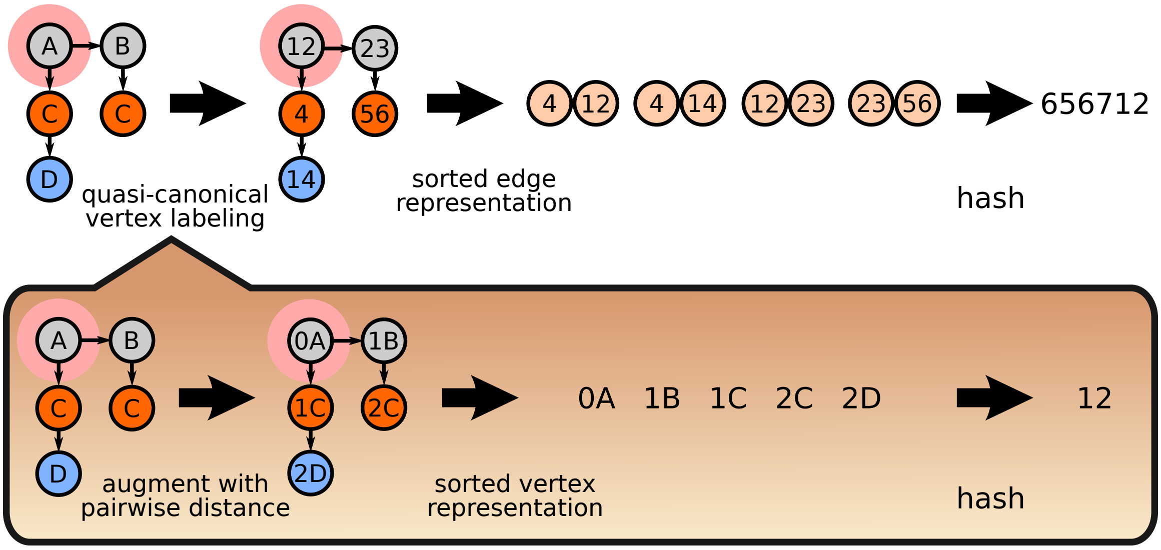

A
B
C
C
D
quasi-canonical vertex labeling
12
23
4
56
14
sorted edge representation
4 12 4 14 12 23 23 56
656712
hash
A
B
C
C
D
augment with pairwise distance
0A
1B
1C
2C
2D
sorted vertex representation
0A 1B 1C 2C 2D
12
hash

**Figure 3.4.:** Graph invariant computation for rooted graphs. Top row: an integer code is obtained from the sorted list of edges by hashing technique. Bottom row: a vertex quasi-canonical label is computed. Here, the root vertex A is converted to an integer code 12.

non-isomorphic graphs have same encodings or due to a collision introduced by the hashing technique even though they have different encodings.

In graph invariant computation, the graph encoding $\mathcal{L}^g(G)$ was obtained by defining two label functions: (i) for the vertices ($\mathcal{L}^v$) and (ii) for the edges ($\mathcal{L}^e$). For the vertex $v$, the function $\mathcal{L}^v(v)$ defines a lexicographically sorted sequence, which is a series of pairs composed of a topological distance, and a vertex label, $\mathcal{L}^v(v)$, that returns a sorted series of pairs for all $u \in G$. By composing the original edge label and the new vertex label, the new edge label, $\mathcal{L}^e(uv)$, is produced. Then, the sorted lexicographically series is assigned to the graph $G$ by $\mathcal{L}^g(G)$. Finally, a construction based hashing technique, proposed by Damgård [245], is used to map the variable length data into various lists of integer codes [236]. The graph invariant computation process is depicted in Figure 3.4.

### Advances of graph kernel

In [246], Heyne *et al.* extended the work in [236], introducing two enhancements: direct graph specialization and explicit feature encoding. In this thesis, we have taken the similar strategy; while in the original formulation only undirected graphs are considered, we employed directed graphs as they can better model long biomolecules (such as proteins and DNA) that have a natural direction. To do so, we make sure that the paths used to define the pairwise distances are directed and then we duplicate the input graph. All the edges of the copied graph are reversed and its label dictionary is made non overlapping with the original dictionary. In this way, all the features that are specific for the original and the inverted direction can be created (see Figure 3.5).

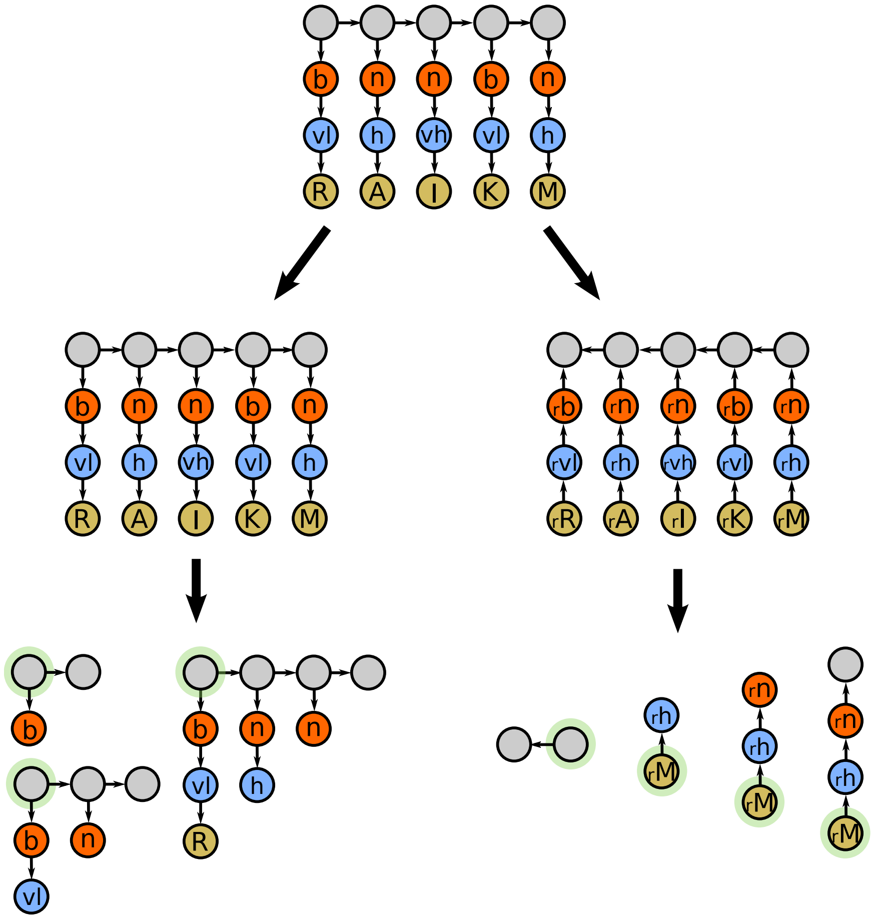


**Figure 3.5.:** Illustration of the direction treatment. Upper row: the original directed graph. Middle row: the original directed graph duplicates into two copies and one copy is used for the inverse direction. Bottom row: the `NSPDK` features for $r = 1, 2, 3$ and $d = 0$ . Root vertices are highlighted. Note that the features are direction specific, since different label alphabets are used for the inverse direction.

Finally, instead of returning only the kernel score, we exposed the hash code of each feature and its associated value as a sparse vector of high dimensionality (see Figure 3.6). This allows a flexible manipulation of the resulting instances and the possibility to use fast stochastic gradient descent methods for model's parameter estimation.

### 3.3.3. Negative class definition

High-throughout experiments often provide only positive interaction data (e.g., phage display) and no or less information regarding negative interaction data. Thus, one of the main problems of most of the machine learning approaches to predict the binding partners of modular protein domain is to generate confidence negative data (see Section 2.3.2 for more details). To tackle this problem, we have developed two different models: (i) a generative model and (ii) a combined model based on one-class and semi-supervised methods. These models are describe below.

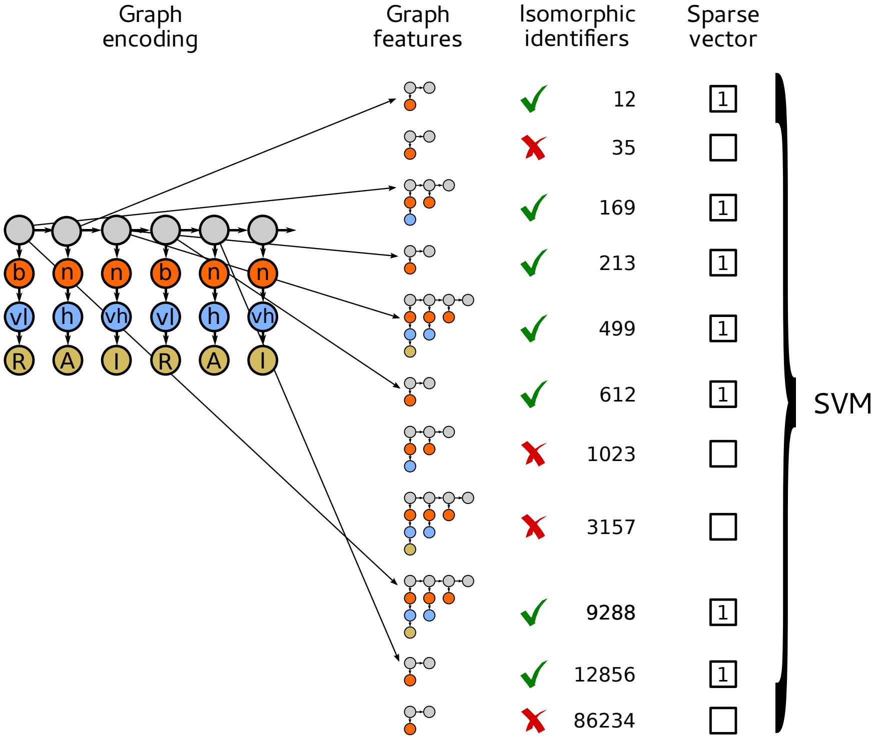


**Figure 3.6.:** Illustration of generating sparse vector from a graph. An isomorphic identifier is assigned to each graph feature and based on their presence, a sparse vector is generated. Finally, the sparse vector is used by the SVM.

### Generative model

The key idea here is to employ a generative approach to model each peptide class and select a subset of instances that is not recognized by any specialized model. We take an approach similar to [199] and select confidence negative interactions using profile based models (i.e., PWMs). In order to better represent the binding specificity of each domain, instead of using a single model, we resort to multiple PWMs, namely one for each motif class for each SH3 domain.

In more detail, we first used the `Fuzzpro` pattern search program from the `EMBOSS` package [247], which uses optimum searching algorithm for finding the pattern and can be used for searching the exact pattern or various ambiguities of the sequences. By using `Fuzzpro`, we clustered the peptides into eight groups, one for each known motif class. We found that the majority of the peptides belong to the canonical motifs of class I and/or class II, while the rest belong to atypical motifs, mainly PxRP, PxxxPR, PxxDY, RxxKP motifs (see Table A.2.1). Afterwards, we used the popular EM-based `MEME` [248] algorithm to generate a PWM, which describes the probability of each possible amino acid in each position in the sequence, for each group.

Finally, we used `MAST` [249], a sequence homology search algorithm, to identify the peptides matching the `MEME` generated PWMs. `MAST` is an efficient tool and was successfully applied for searching DNA motifs in transcription factor binding sites [250]. `MAST` ranks the input sequences according to an E-value type of score. We consider the peptides with a high E-value (i.e., those that are not recognized with confidence by the model) as negatives. The cut-off score was set to the maximum E-value calculated for the known positive instances. Finally, for each domain, we select those peptides that are not recognized by any of the class specific PWMs. By doing so, we identify a total of ∼200K (262883) negative interactions for the whole set of 70 human SH3 domains (see Table 4.4). Note that peptides considered as negative but that are close to the cut-off-score are in fact structurally quite similar to positive peptides.

Training and testing a model using only high-confidence negative interactions can in principle induce a bias. To rule out such a case, we perform an additional experiment (see Section 4.3.5) where we do not filter in any way the negative data.

#### One-class semi-supervised model

The key idea here is to use the SVM one-class approach, pioneered by [251], to warm-start the self-training method for semi-supervised learning [252], restricting the prediction to negative instances only. In [251], it is shown how, in order to identify a region that contains with a high probability most of the positive data, one can formulate the classic SVM optimization problem for binary classification using the origin of the feature space as the only negative instance. In the case of normalized kernels, this boils down to using negative instances that are just the symmetric counterparts of the available positive instances. Here, we follow this latter way given that we can produce the explicit sparse encoding and therefore can efficiently invert each instance.

The self-training approach to semi-supervised learning [252] is a wrapper method that iteratively uses the class predictions over the unlabeled data as true labels for a successive training phase until convergence to a stable state is reached. Here, we use the one-class model to initially induce the class information on the unsupervised instances, but, rather than using both positive and negative predictions, we accept only negative predictions. We select those instances that are predicted with the highest confidence (i.e., that are further away from the class boundary hyperplane) and use them to iteratively train the SVM model. For simplicity, we fix the fraction of the accepted negatives to 50% of the total number of unsupervised instances.

### 3.3.4. Modeling with graph kernel features

Our approach is based on a graph encoding that allows to model relations between specific amino acids as well as different amino acid abstractions. This graph is then processed by the Neighborhood Subgraph Pairwise Distance Kernel (`NSPDK`) [236], that extracts as explicit features, the occurrence counts of all the possible pairs of near small neighborhood subgraphs. The subgraph pairs are characterized by a radius and by a topological distance parameter (for details see in Section 3.3.2). The final classification task is then performed by a support vector machine (SVM) based on the NSPDK. Note that by using an explicit vector encoding, we gain efficiency since we avoid to compute and store the pairwise similarity matrix.

### Single Domain Modeling

When developing models for each specific domain, we only encode information on the candidate peptide sequence as described in Section 3.3.1. Different values for the radius parameter give rise to the parts illustrated in Figure 3.3.

Given the directed nature of the encoding graph, each neighborhood subgraph includes only amino acid that are downstream w.r.t. the current root node. With radius 1 and distance 0, each labeled vertex is considered independently: the corresponding feature representation encodes the frequency of each physico-chemical property (either the charge, the hydrophobicity or the amino acid type) in the single peptide; radius 2 allows properties of adjacent residues (e.g., hydrophobicity and adjacent charge information) to be modeled; radius 3 allows all properties for a single residue to be taken into account jointly. Even larger radius values can capture the joint information for adjacent pairs, triplets etc, of residues. When pairs of neighborhood subgraphs at different distances are used, the composition of the sub-sequence between the two root vertices is ignored allowing a *don't care* or *soft* type of feature matching. Note that the order in which the properties are encoded is chosen so to avoid generating features that subsume each other (i.e., given a neutrally charged amino acid, one can have multiple values for the hydrophobicity but not the other way around). The final descriptors for each peptide contain all features with radii ranging from 0 up to $R_{max}$ and distances in $[0, D_{max}]$. The optimal ranges are determined experimentally via cross-validation techniques. Finally, the training phase allows the determination of the weight distribution on all feature types (general and specific) to obtain optimal predictive performance.

### Multiple Domains Modeling

Several SH3 domains in the human genome bind strongly with class I and/or class II peptides. SH3 domains for FYN, BTK, HCK, FGR, SRC, and LYN proteins are among them. The intuition underlying the multiple domains approach is that, if we are able to exploit the similarities across these domains, we can then increase the predictive performances for each specific domain. In practice, we would be performing a form of *transfer-learning* [253] from one protein domain to another, so that the examples used to induce a model on one domain would also contribute to form the bias of related models, increasing the effective number of available training instances.

To do so, we proceed by coupling the peptide information with the encoding for the domain in a joint feature space; more specifically, we encode the domain amino acid sequence information via its projection w.r.t. the domain consensus alignment. Here, the backbone vertex labels encode the specific position of the amino acid within the reference alignment. By introducing these absolute reference ids, all features (those describing physico-chemical properties and those describing the amino acid composition) become position specific. This absolute reference creates a joint feature space that ultimately allows information about interactions with different domains to be shared.

Note that we are not trying to model the exact pairs of interacting amino acid residues (one in the peptide and one in the domain), as done in [205]. To do so, would imply resorting to resolved protein complexes information, which is not available in large scale. Rather, we represent the candidate interacting peptide and domain as a pair of disconnected graphs. The NSPDK procedure alone, does not instantiate features that can directly express the relationship between parts of the peptide and of the domain sequences. However, we can take full advantage of the *kernel trick* and employ non-linear (i.e., polynomial or Gaussian) kernels. By doing so, the peptide-domain complex is implicitly represented by features that express combinations of the original features. We then rely on the statistical analysis of high-throughput experiments to infer the importance of each position specific features in the domain combined with non-position specific features of the peptide sequence.

## 3.4. PDZ modeling

In this section, we present a Gaussian kernel-based efficient machine learning method for the prediction of PDZ domain mediated interactions. Here, we show that the domain coverage can be increased by applying an accurate clustering technique and thus all the PDZ domains are clustered based on their binding specificity. The prediction models are available across several species (i.e., human, mouse, fly, and worm). Class imbalance problem and generating confidence negative data are handled by the semi-supervised learning technique. Finally, we build models that allow higher-order dependencies between the amino acids in the binding peptides. Additionally, we also show how to build the models using PDZ-peptide complex structure information.

### 3.4.1. Clustering of PDZ domains

We clustered all the available PDZ domains using Markov clustering algorithm (MCL) based on their global sequence identity [254]. MCL is a popular and efficient method for clustering biological sequences and was successfully applied for clustering of protein families [255]. Recently, Li *et al.* have proposed that PDZ domain pairs with greater than 50% sequence identity share similar binding specificity [200]. Thus, we defined 50% sequence identity as a cut-off value to represent similar specificity. We used Needleman-Wunsch algorithm in order to calculate pairwise sequence identity of all PDZ domains. PDZ domain pairs with less than 50% sequence identity were discarded to reduce the noise [256]. In the MCL method, PDZ domain sequence identities can be considered as a weighted graph, where the domains are the nodes and the identity relationships are the edges. Since the MCL algorithm was specifically designed for simple and weighted graphs, clustering of the PDZ domains using MCL is highly reliable. We applied MCL algorithm with 1.4 inflation parameter. This parameter was used for controlling the granularity or the tightness of the clusters and we found 1.4 as the best inflation value for clustering of PDZ domains. Only families with at least two PDZ domain sequences were considered. Finally, 515 PDZ domains from human, mouse, fly, and worm were classified into 138 different families.

### 3.4.2. Feature encoding

Previous studies showed that the C-terminal residues of a peptide are the most important for PDZ-peptide binding specificity [92]. We followed the literature and restricted the peptide sequence to 5 C-terminal positions, namely we extracted the amino acids in positions from P0 to P$-4$ downstream where P0 is the last C-terminal position. We have developed two types of feature encoding methods: (i) sequence-based and (ii) contact-based feature encoding.

In the sequence-based feature encoding, a peptide sequence was mapped into a binary vector $x$, living in a $20 \times 5 = 100$ dimensional space. I.e., for each position, we reserved 20 dimensions (one for each amino acid type) and encoded the amino acid type with a 1 in the corresponding dimension and 0 elsewhere.

For the contact-based feature encoding, we used an approach similar to the one described by Chen *et al.* [88]. Here, the important position pairs (one amino acid from the domain and another from the peptide) were taken into account. First, we constructed a cluster-based, PDZ domain, multiple sequence alignments using MAFFT [195]. We then considered the core position pairs that are in close proximity and extracted only the position pairs with distance less than 4.5 angstroms using domain-peptide complex structures. Note that we have used different reference structures for different families. Each position pair was encoded as a binary vector $x$, living in a $20 \times 20 = 400$ dimensional space. All the position pairs were then encoded in a binary vector of size $400 \times n$, where $n$ is the number of binding pairs. Finally, the sequence-based encoding was concatenated with the contact-based encoding, which produced a binary vector of size $100 + 400 \times n$.

For each domain $D$, a dataset encoded as a set of pairs $(x_1,c_1)$,..,$(x_n,c_n)$ have been compiled where, $x_i$ is the binary feature vector for peptide $P_i$ with the class label $c_i \in \{-1, 1\}$. The class label is $+1$ if the domain $D$ interacts with peptide $P_i$ and -1 otherwise.

### 3.4.3. Predictive model

We employed regularized Gaussian kernel support vector machines to build predictive models [231]. Here, the Gaussian kernel is more suitable than the polynomial kernel, since more data was available for PDZ-peptide interactions and thus performed better. Moreover, this kernel allows the infinite feature dependency. SVM$^{light}$ software was used to build the SVMs [232].

### Gaussian kernel

Gaussian kernel is one of the successful and extensively studied kernel functions used in support vector machine. It forms the hidden units of a Radial basis function networks (RBF), and hence it is also known as the RBF kernel [257]. The Gaussian kernel produces a graph of a characterized symmetric bell curve shape. For a feature space with two inputs (vectors) $x_1$ and $x'$, and a $\sigma > 0$, the Gaussian kernel can be defined as:

$$K(x, x') = \exp\left(-\frac{\|x - x'\|^2}{2\sigma^2}\right), \tag{3.12}$$

where $\|x - x'\|^2$ represents the squared Euclidean distance between two vectors $x_1$ and $x'$. Note that we do not restrict ourselves for using the Euclidean distance in the input space [257]. For example, if $K_1(x, x')$ is a kernel corresponding to a feature mapping $\phi_1$ into a feature space $F_1$, the $\|\phi(x) - \phi(x')\|^2$ can be represented as:

$$\|\phi(x) - \phi(x')\|^2 = K_1(x, x) - 2K_1(x, x') + K_1(x', x'). \tag{3.13}$$

Hence, the Gaussian kernel function can then be defined as:

$$K(x, x') = \exp\left(-\frac{K_1(x, x) - 2K_1(x, x') + K_1(x', x')}{2\sigma^2}\right). \tag{3.14}$$

The parameter $\sigma$ allows the flexibility for the Gaussian kernel, similarly like $d$ parameter in polynomial kernel. A small value of $\sigma$ is equivalent to a large value of $d$. A small value of $\sigma$ allows a classifier to fit any labels and therefore caused an overfitting situation. Conversely, a large value of $\sigma$ will make the function almost impossible to learn any non-trivial classifier, since it gradually reduce the kernel to a constant function. Although, for every values of $\sigma$, the feature space has infinite dimensions, however, the weight decays very fast on higher-order features, if the large values of $\sigma$ are used [257]. In Equation 3.12, if $\frac{1}{2\sigma^2}$ is replaced by a simpler parameter $\gamma$ then the kernel can be defined as:

$$K(x, x') = \exp(-\gamma\|x - x'\|^2). \tag{3.15}$$

Another important parameter in Gaussian kernel is the cost parameter ($C$), which works same way as described in Section 3.2.2.

## 3.4.4. Negative class definition

Datasets derived from high-throughput experiments usually suffer from a lack of reliable negative interaction data. In our study, we were only able to obtain the negative interaction data from a microarray experiment, although the dataset had an imbalance problem. Other data sources (i.e., phage display and PDZBase) provide only positive interaction data. Previous studies showed that machine learning methods work poorly when the dataset is highly imbalanced [216, 217, 235]. In order to generate more negative data, we have employed a semi-supervised learning approach (SSL). The general strategy of SSL is to learn from a small amount of labeled data and a large amount of unlabeled data. Here, differently from the general problem formulation for SSL, we were interested in using the unsupervised material to have a better characterization only of the minority class; in our case, the negative class. Albeit, there are several strategy to deal with SSL problem, we have chosen the self-raining approach and proceeded in an analogous fashion described in Section 3.2.3. Since dataset I was only comprising of mouse PDZ-peptide interaction data, we used all the C-terminal peptides from mouse proteome as unlabeled data.

Finally, the predicted unlabeled peptides having the probability of 0.5 to 0.8 towards the negative class were considered. We ignored very high scoring (probability more than 0.8) predicted negative peptides since they might be very far from positive class, and therefore could produce low quality models. We randomly chose negative data from the pool of predictive negatives, added them to the training data, and re-trained the classifier. In general, there was five times more negative data than positive data, which is computationally feasible [191].

Note that we need both positively and negatively labeled data to apply the described SSL approach, since we need to train the base classifier with both positive and negative data. In our study, we could only employ the SSL approach to domains that occur in dataset I as it contains both classes. For those PDZ domains, where only the positive data was available, we chose the negative data randomly from C-terminal peptides of the respective organism from `UniProtKB/Swiss-Prot` [19]. Note that we only used the negative interaction data from the semi-supervised learning for the training sets, while our test sets contained only experimentally verified positive and negative interaction data.

### 3.4.5. Model fitting protocol

The model hyper-parameters (i.e., $\gamma$ and the cost parameter $C$) were chosen by using 5-fold grid search method to trade-off generalization for data fitting.

We used a 5-fold *stratified* cross-validation in order to evaluate the predictive performance of each model. Here, the data is partitioned into 5 parts ensuring the same proportional distribution of positive and negative instances in each part (see Section 3.6.1 for details). In the cross-validation step, only the families with at least 10 positive data and 10 negative data were taken into account so that each test set contains at least 2 positive and 2 negative interactions. The predictive performance was achieved by using different statistical measures (see Section 3.6.2 for details) under a *stratified* 5-fold cross-validation scheme.

## 3.5. Key advantages of the proposed methods

In Chapter 2, we have discussed several computational methods that have been developed to predict the modular domain mediated interactions. We have also discussed the major drawbacks of these methods. In our study, we have tried to circumvent the limitations of these existing computational methods. The strategies that we have taken to tackle the limitations are described in this section.

### 3.5.1. Non-linear model

Positional dependencies between the amino acids in the binding peptides play an important role to describe the binding specificities of modular domains [43, 188, 191]. PWM-based models (e.g., `Scansite`, `SMALI`, and `MDSM`) and linear machine learning models (e.g., `DomPep`) completely ignore the higher-order dependencies between the amino acids and thus failed to explain accurate binding specificities of the modular domains (see Section 2.3.1 for details). To overcome the linearity problem, we built non-linear models by using appropriate kernel functions that allow higher-order dependencies between the features for all domains (i.e., SH2, SH3, and PDZ). Detailed information about the kernel functions and non-linear models can be found in Section 3.2.2, 3.3.2, and 3.4.3.

### 3.5.2. Data balancing and confidence negative data refinement

It is already known that severe data imbalance problem negatively affect the predictive performances of machine learning methods [210]. Fortunately, in our case, we could circumvent the data imbalance problem by exploiting a useful property of the datasets we had at our disposal: instead of creating novel instances, we could make use of a large quantity of results available from high density peptide array experiments; specifically, we selected those peptides for which no definitive interaction information was available. In this way, we did not have to invent plausible biological peptide sequences to populate the neighborhood of minority class representatives. Rather, we had to perform the easier task of estimating when an existing peptide is likely to belong to the minority concept.

Data derived from high-throughout techniques often suffer from a lack of reliable negative data (see Section 2.3.2), and it is already known that high-confidence negative data is important to build a good machine learning model [216, 217]. In our study, we viewed this whole problem as semi-supervised learning task (SSL). Although there are several strategies available to deal with SSL problem, we have resorted to the self-training strategy in order to get high-confidence negative data, because it was the most suited approach for our methods. See Section 2.3.2 for details about semi-supervised learning problem. The self-training approach is a straightforward yet effective wrapper technique that can be applied to any classifier. It consists an iterative procedure where at each stage the current model predicts the class label over the unsupervised material. In the next training phase, the class labels for the most confident predictions are used. The procedure can then be iterated. Note that this strategy is only applicable when at least a few instances for both classes (positive and negative) are present. In our case, we could use this self-training strategy for SH2 and PDZ domains (see Section 3.2.3 and 3.4.4) but not for SH3 domains, since we had only positive interaction data for SH3 domains. However, for SH3 domains, we also tried to use one-class semi supervised technique to achieve high-confidence negative data but unfortunately, we achieved under-performing models. To get high-confidence negative data for SH3 domains, we developed a generative approach based on multiple PWMs to model each peptide class for a specific SH3 domain (see Section 3.3.3). We retrieved a subset of peptides that were not recognized by any PWMs and considered them as reliable negative instances. These negative instances along with true positive instances were then used for training the binary classifier.

### 3.5.3. Alignment-free approach

An initial multiple alignment of the binding peptides is necessary in existing PWM-based methods, but this is a hard task for proline-rich SH3-bound peptides. Even minor alignment errors typically introduce significant noise in PWMs estimate. For prediction of SH3 domain mediated interactions, a PWM-based tool, namely `MUSI`, has been developed recently, which can recognize multiple specificities of SH3 domains, but severely affected by error-prone peptide alignment (see Section 2.3.1 for details). In an other recent publication, the authors have tried to tackle the peptide alignment problem by performing two essential tasks simultaneously: alignment and clustering using Gibbs sampling approach and identifying biologically relevant binding motifs that cannot be described well with a single PWM [258]. However, this approach cannot fully circumvent this problem since they anyway rely on an alignment. Our approach sidesteps these issues all together, as we have completely eliminated the need for an initial peptide alignment in the case of SH3 domains where the optimal alignment of peptides is almost not possible (see Section 3.3.2 for details).

### 3.5.4. Domain coverage

Most of the available tools suffer from the domain coverage problem. In Section 2.3.3, we have discussed this problem in details. To increase the domain coverage, we combined the interaction data for domains that are similar in substrate specificity, and built a multi-domain model for a specific domain family. In SH3-peptide interaction, we have shown that the multi-domain model has better prediction accuracy than a single domain model (see Section 3.3.4 and Figure 4.6). We effectively applied this strategy in our study and clustered all PDZ domains from human, mouse, fly, and worm based on their specificity (see Section 3.4.1). In this way, we could build a single classifier for similar domains across the organisms.

### 3.5.5. Regularization technique

Machine learning methods are often affected by overfitting problem (see Section 2.3.2). In our study, we have successfully handled this problem. Among the several ways to ensure a regularized solution, we adopted the strategy championed in SVM, i.e., we minimized the complexity of the model by constraining the size of $w$ and the hyper-parameters, e.g., the degree of the polynomial $d$. We did this using a cross-validation procedure in order to achieve a good compromise with respect to the training misclassification error. In practice, the SVM optimal hyperplane is determined as the solution to a minimization problem where the objective function combines a term proportional to the training error and a term proportional to the complexity of the model (computed as the norm of the hyperplane coefficient vector). The mixing coefficient that weights the importance of the error w.r.t. the model complexity and the hyper-parameters were selected from a finite set of alternatives. The best parameter combination was chosen by evaluating the predictive performance of each specific model over a held out set of instances (the validation set). Note that the performance of the selected model was evaluated over a further held out set of instances (the test set) that had never been used neither in the training phase nor in the validation phase.

## 3.6. Predictive performance estimations

### 3.6.1. Cross-validation

For the evaluation of machine learning methods, cross-validation technique is more commonly used. More specifically, this technique is used for estimating how accurate a model will perform in practice. In the cross-validation setup, the dataset is randomly split into $n$ equal parts where *(n − 1)* used for training and remaining data is used for validation. This process is then repeated for $n$ times. 10-fold or 5-fold cross-validation is a standard way to measure the error rate of a predictive model, however, other types of cross-validation techniques are also successfully used. One such prevalent technique is leave-one-out cross-validation, which is also an $n$-fold cross-validation, where $n$ is the number of instances of the dataset. These cross-validation techniques are depicted in Figure 3.7.

A stratification procedure is used to maintain approximately the same proportion of the two types of class labels (positive and negative) in each fold. Cross-validation with stratification procedure is known as *stratified* cross-validation.

Final performance is calculated by averaging all the predictions from all runs. The predictive performance of each problem is measured by different statistical measures (e.g., accuracy, recall etc.).

### 3.6.2. Performance measure

We formulated a learning problem for each modular domain and/or domain family. The predictive performance for each problem was mainly assessed by computing 5 measures: sensitivity, specificity, precision, area under the receiver operating characteristics curve (AUC ROC), and area under the precision recall curve (AUR PR).

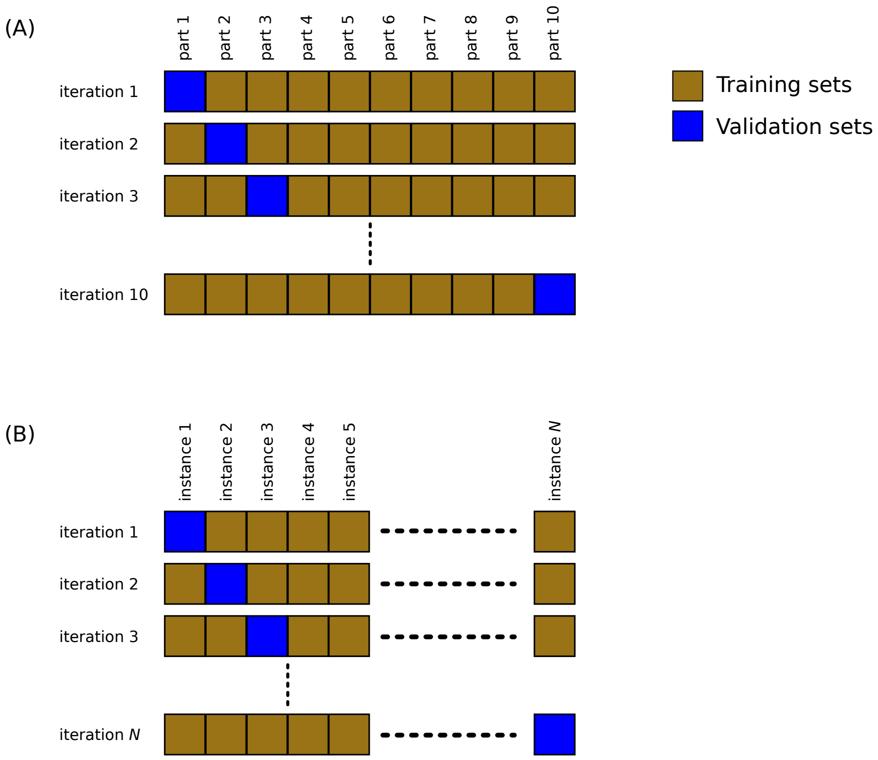


**Figure 3.7.:** (A) 10-fold cross-validation: In this setting, $N$ instances are split into 10 equal parts. 9/10 of the data are used for training and remaining part used for validation. (B) leave-one-out method: In this setting, $N$ instances are split into $n$ parts where $N = n$.

**Sensitivity:** The sensitivity provides the proportion of actual positive instances that are correctly identified as positive. Sensitivity is also known as Recall:

$$Sensitivity/Recall = \frac{TP}{TP + FN}. \tag{3.16}$$

**Specificity:** The specificity provides the proportion of actual negative instances that are correctly identified as negative:

$$Specificity = \frac{TN}{TN + FP}. \tag{3.17}$$

**Precision:** The precision provides the proportion of identified instances that are actually positive:

$$Precision = \frac{TP}{TP + FP}. \tag{3.18}$$

Here, TP denotes true positive, FP denotes false positive, TN denotes true negative and FN denotes false negative, which are the possible outcomes of a two-class prediction (see Table 3.1).

In our case, TP means true domain-peptide interactions are correctly predicted as binding interactions; FP means true non-binding interactions are incorrectly predicted as binding interactions; TN means true non-binding interactions are correctly predicted as non-binding interactions; and FN means true binding interactions are incorrectly predicted as non-binding interactions.

**Table 3.1.:** Possible outcomes of a two-class prediction.

| | | Prediction class | | |
|---|---|---|---|---|
| | | Positive ($P$) | Negative ($N$) | Total |
| Actual class | Positive ($P$) | True Positive ($TP$) | False Negative ($FN$) | $TP + FN$ |
| | Negative ($N$) | False Positive ($FP$) | True Negative ($TN$) | $FP + TN$ |
| | Total | $TP + FP$ | $FN + TN$ | $N$ |

**AUC ROC:** The area under the receiver operating characteristics curve (AUC ROC) is commonly used for assessing the trade-off between hit rate and false alarm rate over a noisy channel [207]. In simple terms, the AUC ROC is obtained by plotting the fraction of true positives out of the positives (TPR = true positive rate) vs. the fraction of false positives out of the negatives (FPR = false positive rate), at various threshold settings. Note that AUC ROC shows the performance of a classifier without considering the class distribution or cost errors [207].

**AUC PR:** The area under the precision recall curve (AUC PR) is defined as the area under the curve obtained by plotting precision as a function of recall.

# Chapter 4

# Applications and performance evaluations

## 4.1. Overview

In this chapter, we describe the application and performance evaluation of all three methods that have been presented in Chapter 3. This chapter is divided into four sections; first three sections report the dataset compilation and the application of the three proposed methods. We show that our methods achieved significantly better predictive performances with respect to state-of-the-art approaches. Furthermore, we have tested our methods on manually curated and reliable datasets, and also achieved better performances than other existing methods. In the last section of this chapter, we have performed a genome-wide prediction for SH2, SH3, and PDZ domains to uncover novel modular domain-peptide interactions that have important biological insights. Additionally, we have performed a term-centric analysis, which identifies enriched biological annotation terms (GO-term, pathways etc.) associated with input protein list, for the top interactions predicted by our tools to unveil novel functionalities of the interactions. Finally, we make all the top genome-wide predictions freely available to the scientific community. The work presented in this chapter is a part of the following publications: [P2], [P3], and [P4].

## 4.2. SH2-peptide interactions

### 4.2.1. Introduction

Previous research showed that the dependencies between different ligand positions take important role in the binding specificity of the SH2 domains [43]. In recent years, polynomial kernels have been successfully applied to the prediction of DNA-protein interactions [259]. We used domain specific non-linear models for SH2-peptide interactions that are based on support vector machines (see Section 3.2.2 for details). As the complexity of the model increases, so does the required number of training instances. While modern high-throughput techniques seem to be the perfect solution to the data requirements, they have other issues. The first problem is that the techniques, such as pool oriented peptide arrays [39, 156] do not test individual peptides but pools of peptides with common properties. In the second phase, individual peptides are tested with separate methods. Thus, while these approaches provide information about real interactions (positive data), they cannot be reliably used to assess the lack of a domain-peptide interaction. A similar situation occurs with many high-density peptide arrays where affinities are not reported. Other high-throughput approaches like microarrays do report affinities (e.g., [44] and [165]) and thus can be used to assess the lack of strong interaction. However, these approaches suffer from a low signal to noise ratio and produce results that are often inconsistent. For example, in one microarray experiment [165] found that the number of interactions between 11 peptide sequences extracted from protein ErbB1 and 85 SH2 domains is 37, while under similar settings in another microarray experiment [44] found three times as many interactions.

This state of affairs leads to a great imbalance between the available information on positive vs. negative interaction data. Such an imbalance constitutes a severe problem when fitting a predictive model. For example, for some SH2 domains, the positive interaction data can be up to 15 times more abundant than the negative interaction data. It is known that in these conditions predictive systems produce suboptimal results. However, in this thesis, this data imbalance problem is successfully handled by a semi-supervised iterative approach as described in Section 3.2.3. Finally, we devise non-linear support vector machine (SVM) models for 51 human SH2 domains.

We show that our approach performs significantly better than state-of-the art SH2-peptide interaction prediction tools. Furthermore, when applying it on high quality hand-curated SH2-peptide interaction data from `PhosphoELM` database [260], we achieved higher True Positive Rate (TPR) in comparison to PSSM models (`SMALI`) and energy model. In addition, we perform a genome-wide analysis and find interesting insights of biological relevance.

### 4.2.2. Dataset compilation

#### Dataset I (high density peptide array data)

From the NetPhorest database [191], we collected information on 61 SH2 domains and 920 phosphorylated peptides for a total of 14678 interactions. After removing all redundancies, we obtained 7544 positive interactions.

Note that for high density peptide array experiments, there is evidence only for positive interactions. One cannot, however, assume that the remaining $(61 \times 920) - 7544 = 48576$ interactions are of the non-binding type (i.e., negative interactions). It can happen that these domain-peptide interactions were just not observed in the assay due to the experimental stringency (e.g., consistency among replicates).

#### Dataset II (microarray data)

From the protein microarray experiments in [165], we have considered the SH2-peptide interactions data excluding the PTB-peptide interactions. There are 115 SH2 domains and 20 singly phosphorylated peptides from ErbB2, ErbB3, and ErbB4 proteins. Note that there are 10 cases where a single protein has both a C-terminal and N-terminal SH2 domain. Since the database does not report the assignment of which peptide specifically binds to which of the two domains (N and C terminal), we have discarded the interactions related to these proteins. From this dataset, we have collected $105 \times 20 = 2100$ interactions, with 160 positive interactions and the remaining $2100 - 160 = 1940$ being considered as negative interactions.

#### Dataset III (microarray data)

From the protein microarray experiments in [44], we have considered the SH2-peptide interactions data excluding the PTB-peptide interactions. In this study, there are 85 SH2 domains and 41 singly phosphorylated peptides from EGFR, FGFR, and IG1FR proteins. We have proceeded in an analogous fashion as with dataset II, and we have collected $85 \times 41 = 3485$ interactions with 314 positive interactions and $3485 - 314 = 3171$ negative interactions.

#### Dataset IV (curated data)

From PhosphoELM [260], which is a high-quality manually curated database, we have extracted the interactions for 28 SH2 domains with 339 peptides. This dataset was considered for testing.

We have combined positive and negative data from two microarray datasets (dataset II and dataset III) using the measured apparent equilibrium dissociation constant or affinity constant ($K_D$ value) to determine the class label [44, 165]. SH2-peptides interactions with

K$_D$ values lower than 2000 nanomolar (nM) were considered as binding (positive interactions) while all other pairs were considered as non-binding (negative interactions).

The total number of positive interactions was 474 (160 and 314 respectively from dataset II and dataset III), while the total number of negatives interactions was 5111 ($2100 - 160 = 1940$ and $3485 - 314 = 3171$ respectively).

Dataset I contains 7544 positive interactions and no negative interactions. Among the 474 positive interactions in dataset II and III, 247 (112 and 135) were in common between the microarray and the peptide array data. After removing the positive interactions of dataset I from dataset II and III, we obtained 227 (48 and 179) unique positive interactions for dataset II and III.

Surprisingly, we found 149 interactions for which the microarray data and the peptide array data are in disagreement, i.e., it is positive for dataset I but negative for dataset II and III. We have, therefore, discarded those interactions to reduce unreliable and conflicting information in the training phase. As a consequence, the number of negatives from dataset II and III is reduced to $5111 - 149 = 4962$, and the number of positives in dataset I is reduced to $7544 - 149 = 7395$.

To compose our datasets, we used the positive interactions from the dataset I (7395) and the available negative interactions from dataset II and III (4962). The non redundant positive data derived from microarray experiments was kept for validation purposes. For each of the 61 SH2 domain in dataset I, we compile a separate dataset. We discarded 10 domains that have less than 40 positive interactions, since no complex model can be reliably fit. Finally, we have $61 - 10 = 51$ SH2 domains for which we have 6742 positive and 2523 negative interactions. See Table 4.1 for details.

### 4.2.3. Modeling

Our approach takes as input peptide sequences that have been previously aligned, and as it is common in literature, it is based on amino-acid positional features. The alignment phase induces a global position system where the phosphotyrosine residue is given position 0. Differently from most approaches, we propose to model complex non-linear dependencies between the amino-acid positional features.

**Table 4.1.:** Ensemble data from literature and the final data used in this study after compilation. # D is the number of domains, # P is the number of peptides, # I is the number of interactions, # Pos is the number of positive data, # Neg is the number of negative data, and # Ukn is the number of unknown data. The table is taken from [P4].

| Data source | Original Data | | | | | | Selected Data | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | # D | # P | # I | #Pos | #Neg | #Ukn | #D | #P | #I | #Pos | #Neg | #Ukn |
| Dataset I | 61 | 920 | 56120 | 7544 | – | 48576 | 51 | 880 | 44800 | 6742 | – | 38138 |
| Dataset II | 105 | 20 | 2100 | 160 | 1940 | – | 51 | 20 | 1020 | 48 | 851 | – |
| Dataset III | 85 | 41 | 3485 | 314 | 3171 | – | 46 | 41 | 1886 | 179 | 1672 | – |
| Dataset IV | 63 | 359 | – | 878 | – | – | 28 | 197 | – | 339 | – | – |

Previous studies showed that residues in the close vicinity of the phosphotyrosine are highly predictive for SH2 domain-peptide interaction [181, 183, 230]. For example, it is known that the SH2 domain of CRK binds peptides where amino acid Leu or Pro is in position +3, however, the presence of other amino acids (i.e., His, Arg, Ala, and Pro) in position +1 and +2 can inhibit the interaction. [43]. Thus, we followed the literature and restricted the peptide sequence to 6 specific positions, i.e., we extracted all the amino acids ranging from 2 upstream to 4 downstream of the phosphotyrosine residue. A peptide is therefore mapped into a binary vector $x$ living in a $20 \times 6 = 120$ dimensional space (as the central amino acid is always a phosphotyrosine, it is not informative and is not included in the encoding), that is, for each position, we reserved 20 dimensions (one for each amino acid) and encoded the amino acid type with a 1 in the corresponding dimension and 0 elsewhere (see Section 3.2.1 for more details).

For the predictive model, many popular approaches, such as `SMALI` [183], are based on PSSMs. We note that these methods are essentially linear models and cannot therefore model arbitrary functional dependencies between amino acid positions.

Here, we propose three ways to improve over the PSSM models: (i) upgrading the model from linear to non-linear, (ii) making the system more robust using regularization techniques, and (iii) making an effective use of both interaction information (positive examples) and non-interaction information (negative examples) by dealing with the imbalance issues.

Specifically, non linear models allow to express decision rules that can take into consideration complex functional dependencies between amino acid positions. It could be important to differentiate between the situation where we have co-occurrence of two or more amino acids and the situation where one has independent occurrences of the same amino acids in different peptides. For example, consider a case where the presence of amino acid Asn in position +2 alone is not sufficient to guarantee the interaction and neither is the presence of amino acid Lys in position -1. However, if these amino acids are occurring in their respective positions at the same time, then the binding occurs. Note that there can be different instances of this situation, such as two or more amino acids can have a non-additive effect as described in the example, or two or more amino acids can exclude each other etc. In order to model this non-linear dependencies (but at the same time control the complexity of the model), we upgrade to polynomial kernels (see Section 3.2.2 for details). Note that the degree of the polynomial kernel is optimized via cross-validation and hence a simpler linear model can still be chosen for some SH2 domains when it offers better performance.

The second improvement is to employ regularization techniques to avoid overfitting. Although there are many different ways of dealing with this problem, we adopt the strategy that has been championed in support vector machines. The basic idea of regularization is to minimize the complexity of the model by adding a penalty to discount the cumulative size of the parameters. To be more precise, the complexity of the model depends on the degree of the polynomial kernel (since this determines the number of parameters) and on the cumulative size of the parameter vector in the SVM (see Section 3.2.2 for details).

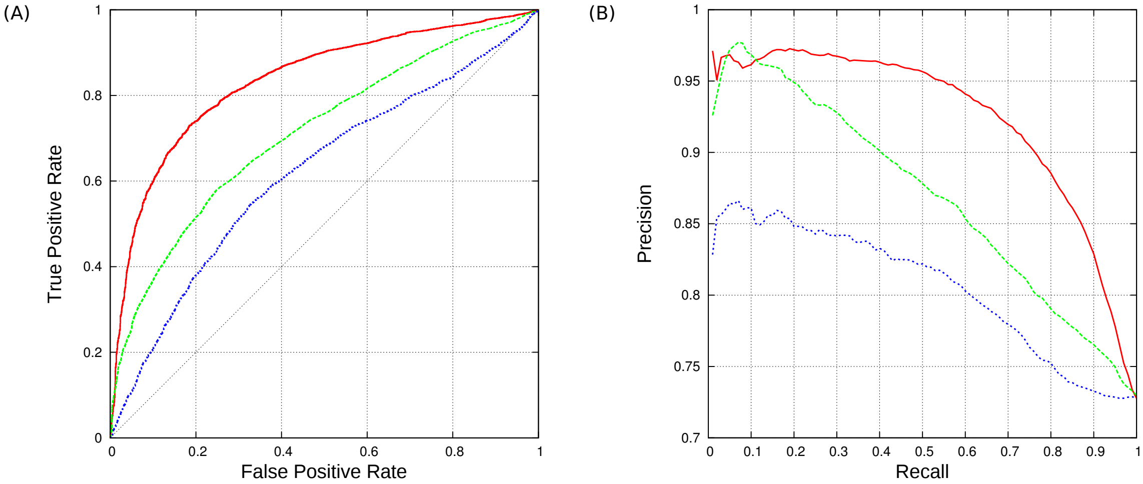


**Figure 4.1.:** Comparison of AUC ROC and precision-recall curve of three different approaches. (A) Showing the comparison of the AUC ROC for the SVM performance (solid red line), the `SMALI` performance (dashed green line), and the performance of energy model (dotted blue line). This figure clearly indicates the SVM performance with 0.83 AUC ROC is significantly higher than the `SMALI` and energy model approaches with 0.71 and 0.62 AUC ROC, respectively. (B) Showing the comparison of the precision-recall curve for the SVM performance (solid red line), the `SMALI` performance (dashed green line), and the performance of energy model (dotted blue line). In this case, the SVM performance with 0.93 precision-recall curve is higher than the `SMALI` and energy model approaches with 0.87 and 0.81 precision-recall curve, respectively. This figure is taken from [P4].

Finally, to tackle the data imbalance problem, we resort to the self-training approach, which relies only on the good discriminative properties of the base classifier. In this approach, the classifier is trained from labeled data, which is then used to test the unlabeled data. The confidence predictions are iteratively added to the training set until the dataset is balanced (see Section 3.2.3 for details).

## Evaluation

In order to assess the expected predictive performance of our approach, we have performed two types of experiments: (i) a cross-validation and random splitting on combined data from three sources: a peptide array library data (dataset I) and two microarray datasets (dataset II and dataset III); and (ii) we performed a validation experiment using a manually curated SH2-peptide interaction dataset (dataset IV) (see Section 4.2.2 for details).

We compare the performance against two state-of-the-art approaches: (i) a tool based on PSSMs, and (ii) an energy model based on interaction maps. The first tool, called `SMALI` [183], is available for 76 SH2 domains and is based on the same peptide representation that we used in our study (i.e., -2 to +4 amino acids with pTyr in $0^{th}$ position). The second tool [230] is an energy model based on different types of interaction maps where only the positions of amino acids found to be in contact are used.

### 4.2.4. Predictive performance evaluation

On each SH2 domain, we evaluate the predictive performance of our approach with a *stratified* 5 fold cross-validation (see Section 3.6.1 for more details about cross-validation settings).

The hyper-parameters, i.e., the polynomial degree, the trade-off between fitting and smoothing cost parameter $C$, are determined using a 10-fold cross-validation on the training set. Using a repeated random split with 75% of the data for training and the remaining 25% for testing, we obtain performance values which are comparable to those obtained in the cross-validation setting (see Figure A.1.1).

We compute the area under the ROC curve (AUC ROC) and the area under the precision and recall curve (AUC PR) (see Figure 4.1). Additionally, in Table 4.2, we report sensitivity and specificity with standard deviation per domain for different treatments of negative data, where the second column refers to no imbalance treatment, the third refers to a random re-balancing strategy, and the last refers to the proposed iterative self-training strategy.

To assess the importance of the dependency between the amino acid positions, we also compared the predictive performance of a linear v.s. a non-linear (i.e., polynomial with degree 2) kernel. In $42/51 = 82.3\%$ cases, the polynomial kernel outperformed the linear kernel according to the AUC ROC measure, which increases to $47/51 = 92.2\%$ cases when we consider the AUC PR measure (see Table A.1.2).

#### Performance comparison

We compare our results with two state-of-the-art tools: `SMALI` [183], and an energy model approach [230]. We apply these tools as well as our approach to all 51 test sets (`SMALI` could be applied to 45 test sets, as it does not have model for the other 6 SH2 domains). Our model achieves an average AUC ROC of 0.83 and average AUC PR of 0.93 (see Figure 4.1), outperforming the other two approaches: `SMALI` achieves AUC ROC of 0.71 and AUC PR of 0.87; the energy model achieves AUC ROC of 0.62 and AUC PR of 0.81. Detailed information on the AUC ROC and AUC PR for each SH2 domain is available in Figure A.1.2 and Figure A.1.3, respectively.

We note that `SMALI` achieves a very high specificity (0.95 on average) in all 45 SH2 domains when the proposed threshold is used (i.e., relative `SMALI` score 1), however, this comes at the expenses of a very poor sensitivity (0.26 on average). See Table 4.3 for details.

In order to directly compare the sensitivities, we identified the threshold for our model so to achieve the same specificity as `SMALI` (and another threshold for the energy model). The advantage of our approach is evident in this setting too, achieving a sensitivity of 0.45 on average against 0.26 for `SMALI` and 0.17 for the energy model.

### 4.2.5. Comparison on validated data

Here, we test our approach with `SMALI` on a manually curated and reliable database of SH2-peptide interactions called `PhosphoELM` [260]. We could not test the energy model, since

**Table 4.2.:** Comparison of specificity and sensitivity. We compare the sensitivity and specificity of each SH2 domain, achieved by using three different datasets (original imbalanced dataset, balanced dataset with randomly chosen negative data, and balanced dataset with good negative data derived by self training process). * The average is computed over all domains except domains indicated with †. The table indicates the datasets generated by the self training strategy perform better. This table is taken from [P4].

| Domains | Original | | Random re-sample | | Neg Semi-supervised | |
|---|---|---|---|---|---|---|
| | Specificity | Sensitivity | Specificity | Sensitivity | Specificity | Sensitivity |
| ABL1 | 0.54 ±0.08 | 0.84 ±0.1 | 0.84 ±0.17 | 0.45 ±0.09 | 0.75 ±0.14 | 0.68 ±0.14 |
| ABL2 | 0.53 ±0.33 | 0.88 ±0.09 | 0.81 ±0.32 | 0.35 ±0.1 | 1 ±0 | 0.55 ±0.17 |
| APS | 0.64 ±0.11 | 0.82 ±0.08 | 0.88 ±0.13 | 0.55 ±0.16 | 0.67 ±0.14 | 0.74 ±0.13 |
| BCAR3 | 0.44 ±0.29 | 0.72 ±0.07 | 0.7 ±0.1 | 0.38 ±0.15 | 0.55 ±0.18 | 0.56 ±0.09 |
| BLK | 0.55 ±0.14 | 0.92 ±0.04 | 0.8 ±0.11 | 0.63 ±0.07 | 0.7 ±0.19 | 0.78 ±0.11 |
| BMX † | 0.74 ±0.05 | 0.79 ±0.09 | – | – | – | – |
| BRDG1 † | 0.76 ±0.11 | 0.82 ±0.08 | – | – | – | – |
| BTK | 0.54 ±0.11 | 0.78 ±0.08 | 0.86 ±0.1 | 0.36 ±0.16 | 0.88 ±0.1 | 0.64 ±0.2 |
| CRK | 0.67 ±0.16 | 0.97 ±0.03 | 0.96 ±0.1 | 0.68 ±0.12 | 0.85 ±0.13 | 0.89 ±0.05 |
| CRKL | 0.63 ±0.17 | 0.92 ±0.05 | 0.96 ±0.09 | 0.71 ±0.13 | 0.94 ±0.09 | 0.8 ±0.12 |
| CTEN | 0.89 ±0.08 | 0.7 ±0.08 | – | – | – | – |
| E105251 | 0.57 ±0.16 | 0.83 ±0.07 | 0.92 ±0.08 | 0.43 ±0.06 | 0.69 ±0.09 | 0.75 ±0.06 |
| E109111 | 0.65 ±0.29 | 0.89 ±0.04 | 0.88 ±0.07 | 0.55 ±0.11 | 0.81 ±0.13 | 0.67 ±0.15 |
| E185634 | 0.8 ±0.11 | 0.99 ±0.03 | 0.95 ±0.11 | 0.54 ±0.2 | 0.9 ±0.14 | 0.86 ±0.05 |
| EAT2 | 0.66 ±0.2 | 0.94 ±0.05 | 0.85 ±0.04 | 0.63 ±0.09 | 0.83 ±0.1 | 0.85 ±0.11 |
| FER † | 0.92 ±0.06 | 0.85 ±0.14 | – | – | 0.95 ±0.05 | 0.69 ±0.12 |
| FES † | 0.92 ±0.08 | 0.82 ±0.11 | – | – | – | – |
| FGR | 0.54 ±0.05 | 0.86 ±0.09 | 0.78 ±0.13 | 0.71 ±0.05 | 0.64 ±0.15 | 0.85 ±0.09 |
| FRK | 0.42 ±0.33 | 0.96 ±0.04 | 0.72 ±0.3 | 0.66 ±0.18 | 0.65 ±0.25 | 0.86 ±0.07 |
| GRAP2 | 0.93 ±0.08 | 0.97 ±0.03 | 0.9 ±0.07 | 0.94 ±0.06 | 0.95 ±0.08 | 0.96 ±0.04 |
| GRB10 | 0.49 ±0.1 | 0.85 ±0.03 | 0.85 ±0.05 | 0.29 ±0.12 | 0.94 ±0.09 | 0.43 ±0.16 |
| GRB14 | 0.48 ±0.23 | 0.9 ±0.03 | 0.84 ±0.1 | 0.47 ±0.11 | 0.6 ±0.18 | 0.7 ±0.13 |
| GRB2 | 0.87 ±0.05 | 0.91 ±0.06 | 0.91 ±0 | 0.91 ±0.06 | 0.93 ±0.04 | 0.9 ±0.06 |
| HCK | 0.55 ±0.25 | 0.91 ±0.04 | 0.82 ±0.13 | 0.5 ±0.09 | 0.79 ±0.21 | 0.75 ±0.08 |
| INPPL1 | 0.64 ±0.06 | 0.82 ±0.07 | 0.84 ±0.15 | 0.45 ±0.07 | 0.69 ±0.16 | 0.8 ±0.07 |
| ITK | 0.71 ±0.22 | 0.85 ±0.06 | 0.91 ±0.1 | 0.53 ±0.09 | 0.95 ±0.06 | 0.72 ±0.11 |
| LCK | 0.55 ±0.09 | 0.87 ±0.07 | 0.88 ±0.05 | 0.5 ±0.07 | 0.7 ±0.09 | 0.73 ±0.08 |
| LCP2 | 0.85 ±0.04 | 0.76 ±0.07 | – | – | – | – |
| LYN | 0.62 ±0.17 | 0.83 ±0.13 | 0.75 ±0.16 | 0.47 ±0.17 | 0.77 ±0.12 | 0.67 ±0.18 |
| MATK | 0.83 ±0.17 | 0.79 ±0.07 | – | – | – | – |
| MIST | 0.3 ±0.45 | 0.94 ±0.04 | 0.9 ±0.22 | 0.41 ±0.1 | 0.5 ±0.5 | 0.77 ±0.07 |
| NCK1 | 0.63 ±0.11 | 0.83 ±0.08 | 0.78 ±0.09 | 0.51 ±0.17 | 0.84 ±0.14 | 0.71 ±0.13 |
| NCK2 | 0.71 ±0.14 | 0.86 ±0.1 | 0.94 ±0.06 | 0.39 ±0.07 | 0.96 ±0.06 | 0.63 ±0.09 |
| PTK6 | 0.52 ±0.14 | 0.89 ±0.09 | 0.93 ±0.07 | 0.42 ±0.09 | 0.78 ±0.19 | 0.68 ±0.1 |
| SH2B | 0.51 ±0.25 | 0.86 ±0.02 | 0.85 ±0.05 | 0.59 ±0.1 | 0.67 ±0.19 | 0.78 ±0.06 |
| SH2D1A | 0.4 ±0.09 | 0.88 ±0.06 | 0.68 ±0.12 | 0.55 ±0.06 | 0.63 ±0.21 | 0.66 ±0.08 |
| SH2D2A | 0.47 ±0.11 | 0.87 ±0.08 | 0.82 ±0.11 | 0.43 ±0.13 | 0.73 ±0.18 | 0.61 ±0.1 |
| SH2D3C † | 0.61 ±0.21 | 0.9 ±0.04 | – | – | – | – |
| SHC1 | 0.53 ±0.19 | 0.83 ±0.05 | 0.92 ±0.04 | 0.42 ±0.28 | 0.69 ±0.17 | 0.71 ±0.12 |
| SHC3 † | 0.71 ±0.04 | 0.79 ±0.08 | – | – | – | – |
| SOCS2 | 0.45 ±0.27 | 0.96 ±0.04 | 0.9 ±0.14 | 0.52 ±0.1 | 0.7 ±0.21 | 0.89 ±0.1 |
| SOCS5 | 0.6 ±0.42 | 0.99 ±0.03 | 0.8 ±0.27 | 0.51 ±0.17 | 0.9 ±0.22 | 0.84 ±0.12 |
| SRC | 0.35 ±0.16 | 0.95 ±0.03 | 0.85 ±0.16 | 0.61 ±0.07 | 0.65 ±0.21 | 0.73 ±0.08 |
| TEC | 0.57 ±0.11 | 0.9 ±0.09 | 0.8 ±0.1 | 0.53 ±0.13 | 0.72 ±0.08 | 0.76 ±0.11 |
| TENC1 | 0.55 ±0.23 | 0.89 ±0.08 | 0.85 ±0.08 | 0.44 ±0.12 | 0.8 ±0.12 | 0.66 ±0.07 |
| TENS1 | 0.58 ±0.23 | 0.87 ±0.09 | 0.87 ±0.05 | 0.49 ±0.12 | 0.77 ±0.15 | 0.78 ±0.11 |
| TNS | 0.57 ±0.12 | 0.87 ±0.05 | 0.73 ±0.13 | 0.68 ±0.03 | 0.7 ±0.09 | 0.83 ±0.04 |
| TXK | 0.47 ±0.1 | 0.86 ±0.07 | 0.82 ±0.09 | 0.53 ±0.17 | 0.65 ±0.12 | 0.74 ±0.11 |
| VAV1 † | 0.86 ±0.12 | 0.88 ±0.04 | – | – | – | – |
| VAV2 † | 0.82 ±0.11 | 0.83 ±0.14 | – | – | – | – |
| YES1 | 0.53 ±0.22 | 0.83 ±0.05 | 0.75 ±0.2 | 0.43 ±0.07 | 0.73 ±0.21 | 0.69 ±0.12 |
| Avg.* | 0.57 | 0.88 | 0.85 | 0.53 | 0.77 | 0.74 |

**Table 4.3.:** Comparison of sensitivity of three different approaches with fixed specificity. Here, we used the fixed specificity that is generated by SMALI program and then we used the same specificity to find the correspondence sensitivity. † SMALI does not have model for these SH2 domains, therefore, we used high specificity for those domains. * The average is computed over all domains except domains indicated with †. This table is taken from [P4].

| Domains | Specificity | SMALI Sensitivity | Energy-model Sensitivity | SVM-model Sensitivity |
|---|---|---|---|---|
| ABL1 | 0.95455 | 0.21023 | 0.03409 | 0.29545 |
| ABL2 | 0.95238 | 0.07500 | 0.02500 | 0.55000 |
| APS | 1.00000 | 0.15441 | 0.10294 | 0.41176 |
| BCAR3 | 0.96226 | 0.05435 | 0.10870 | 0.28261 |
| BLK | 0.90000 | 0.26271 | 0.36441 | 0.52966 |
| BMX | 1.00000 | 0.11250 | 0.01250 | 0.06250 |
| BRDG1 | 1.00000 | 0.00000 | 0.01176 | 0.40000 |
| BTK | 0.96491 | 0.10680 | 0.03883 | 0.36893 |
| CRKL | 1.00000 | 0.26718 | 0.08228 | 0.57595 |
| CRK | 1.00000 | 0.37975 | 0.00000 | 0.64122 |
| CTEN | 0.87500 | 0.53191 | 0.17021 | 0.74468 |
| E105251 | 1.00000 | 0.04965 | 0.04255 | 0.17021 |
| E109111 | 0.98246 | 0.00000 | 0.05941 | 0.40594 |
| E185634 | 1.00000 | 0.27778 | 0.09722 | 0.66667 |
| EAT2 | 0.96610 | 0.31429 | 0.05000 | 0.37857 |
| FER | 0.98333 | 0.56410 | 0.02564 | 0.51282 |
| FES | 0.88333 | 0.67273 | 0.29091 | 0.87273 |
| FGR | 0.88000 | 0.32117 | 0.49270 | 0.52920 |
| FRK | 0.94444 | 0.21212 | 0.17803 | 0.20455 |
| GRAP2 | 0.88136 | 0.96914 | 0.61111 | 0.96914 |
| GRB10 | 0.98113 | 0.13889 | 0.04167 | 0.38889 |
| GRB14 | 0.87931 | 0.28415 | 0.22951 | 0.49180 |
| GRB2 | 0.88889 | 0.90476 | 0.80952 | 0.90476 |
| HCK | 0.89474 | 0.28241 | 0.32870 | 0.51389 |
| INPPL1 | 0.98361 | 0.12295 | 0.04918 | 0.34426 |
| ITK | 0.88372 | 0.30667 | 0.73333 | 0.78667 |
| LCK | 0.96429 | 0.23256 | 0.09302 | 0.43256 |
| LCP2 † | 0.96721 | – | 0.01695 | 0.57627 |
| LYN | 1.00000 | 0.11966 | 0.02564 | 0.01961 |
| MATK | 0.95000 | 0.11321 | 0.13208 | 0.52830 |
| MIST † | 1.00000 | – | 0.19277 | 0.55422 |
| NCK1 | 0.94118 | 0.50459 | 0.29358 | 0.44037 |
| NCK2 | 0.97917 | 0.31683 | 0.04950 | 0.55446 |
| PTK6 | 0.96667 | 0.33824 | 0.00980 | 0.26961 |
| SH2B | 0.96364 | 0.02198 | 0.11538 | 0.42308 |
| SH2D1A | 0.92982 | 0.19162 | 0.07784 | 0.22754 |
| SH2D2A | 0.88333 | 0.33036 | 0.17857 | 0.47321 |
| SH2D3C † | 0.88889 | – | 0.17105 | 0.65789 |
| SHC1 | 0.98039 | 0.24000 | 0.07333 | 0.36000 |
| SHC3 | 1.00000 | 0.15517 | 0.13793 | 0.12069 |
| SOCS2 † | 1.00000 | – | 0.06250 | 0.39583 |
| SOCS5 † | 1.00000 | – | 0.18571 | 0.75714 |
| SRC | 0.97500 | 0.23476 | 0.16159 | 0.27744 |
| TEC | 0.95918 | 0.19018 | 0.28834 | 0.24540 |
| TENC1 | 1.00000 | 0.13990 | 0.02073 | 0.24870 |
| TENS1 † | 1.00000 | – | 0.00813 | 0.14634 |
| TNS | 0.94643 | 0.24876 | 0.07463 | 0.49254 |
| TXK | 0.94545 | 0.16541 | 0.14286 | 0.43609 |
| VAV1 | 0.87500 | 0.35593 | 0.33898 | 0.88136 |
| VAV2 | 0.93878 | 0.22500 | 0.15000 | 0.62500 |
| YES1 | 0.97500 | 0.21101 | 0.08257 | 0.41284 |
| Avg.* | 0.95 | 0.26 | 0.17 | 0.45 |

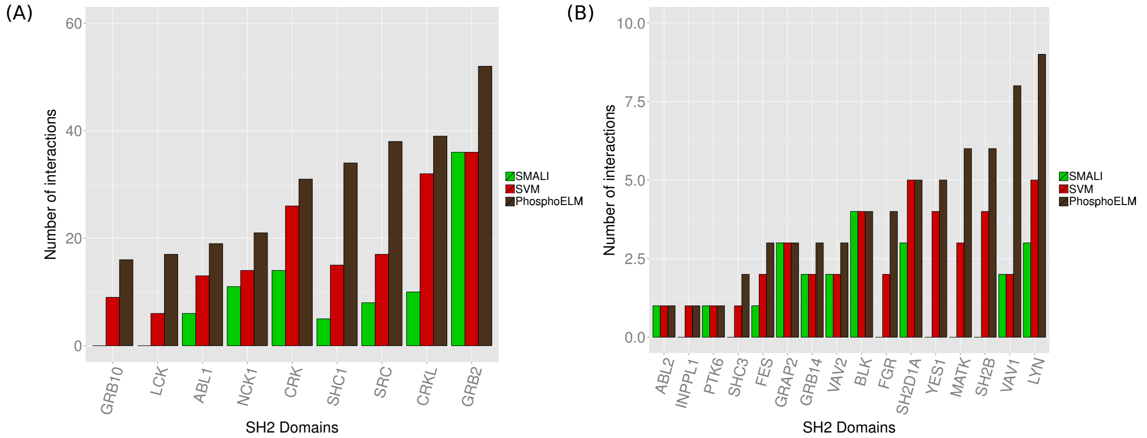


**Figure 4.2.:** Performance evaluation on a manually curated database, PhosphoELM. (A, B) Performance of SMALI and our program on the experimentally validated data. In both (A and B) cases, the brown bars indicate the actual experimentally validated interactions for different SH2 domains, whereas the red and green bars indicate the predicted interactions by SVM models and SMALI, respectively. (A) Showing those SH2 domains that have at least 10 interactions in PhosphoELM 9.0 and (B) showing the SH2 domains that have less than 10 interactions in PhosphoELM 9.0 database. This figure is taken from [P4].

there is no specific threshold that can determine the class. On this dataset the performance of SMALI (comparable to Scansite [181], although with better accuracy for some SH2 domains) is 112 correct interactions predicted over a total of 335 interactions (26 domains, SMALI does not have models for LCP2 and SOCS2 domains), while our approach identifies 213 true interactions (see Figure 4.2). In particular, we correctly predicted all the interactions predicted by the SMALI except two interactions for NCK1 and SRC SH2 domain each.

Note that we have taken care to exclude all the interaction data in the PhosphoELM database from our training sets (unfortunately this cannot be done for the SMALI tool, since we could use only the pre-trained version).

## 4.2.6. Analysis of existing approaches

We further investigate the reliability and the generalization capacity of the two state-of-the-art methods: SMALI and energy model.

### SMALI performance on microarray data

We use dataset II and III to analyze the correlation between the experimental affinity values and the relative SMALI scores. Dataset II contains 3255 interactions between 105 SH2 domains and 31 pY peptides. The strength of the interaction is measured by the apparent dissociation constant [165], denoted as $K_D$. $K_D$ values are available also for dataset III (which contains 3485 interactions between 85 SH2 domains and 41 pY peptides). Interactions are considered reliable when their associated $K_D$ values are lower than 2 $\mu$m.

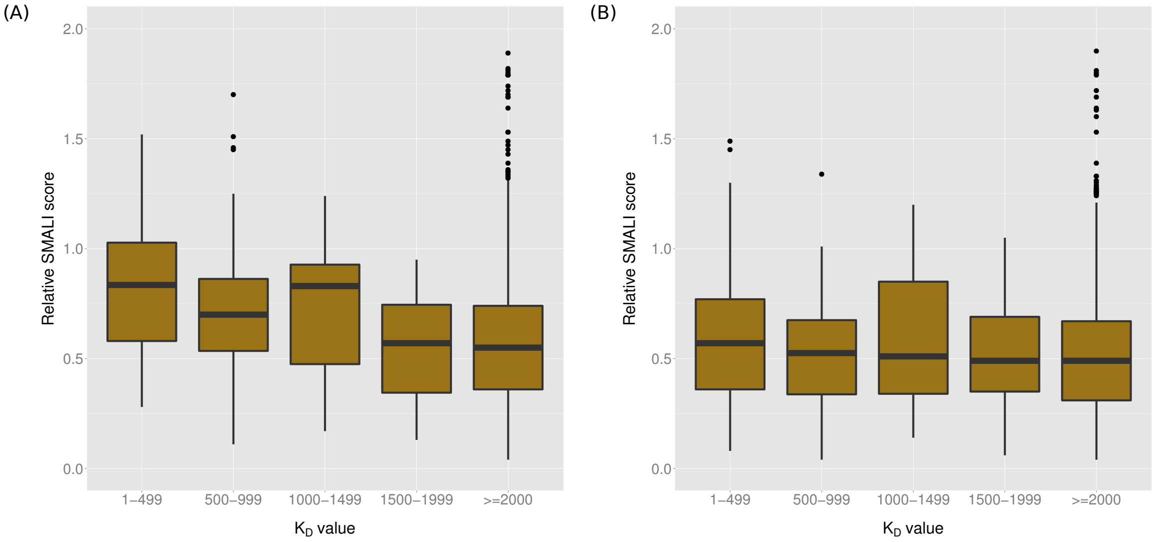


**Figure 4.3.:** Comparison of the relative SMALI scores with two different microarray experiments. (A,B) Barplots of relative SMALI score with microarray experiments. It separates the $K_D$ (apparent equilibrium disassociation constant) into five parts, i.e., 1-499, 500-999, 1000-1499, 1500-1999, and $> = 2000$ (unit is in nm). Among them, $K_D$ values less than 2000 nm were considered as positive interactions, and considered as negative interactions otherwise. (A) Barplot of relative SMALI score with dataset II and (B) barplot of relative SMALI score with dataset III. In both cases, it is clearly observed that there is no correlation between the relative SMALI score and the $K_D$ values. This figure is adopted from [P4].

We compute the relative SMALI score for the SH2-peptide interactions in both dataset II and III. A relative SMALI score $\geq 1$ is considered indicative of a true interaction. In Figure 4.3, we report a box plot for the distribution of the relative SMALI scores vs. the $K_D$ values. We note that a large fraction of interactions that have $K_D$ values lower than 2 $\mu$m (experimental evidence for a strong binding case) have also low relative SMALI scores (no predicted interaction). If we consider only the non binding interactions, we observe a Spearman rank correlation $\rho$ = -0.12 w.r.t. the SMALI score (we would expect a large negative value for good predictive capacity). If we consider the binding interactions, we see that the average SMALI score is 0.53 ± 0.27, significantly below the unit threshold.

An illustrative case is the interaction between domain ABL1 and peptide ErbB2 pY1139, which has an experimentally $K_D$ value of 0.16 $\mu$m (indicating a very high affinity and a high probability of binding) [165]. Here, however, the SMALI tool predicts no interaction, giving a relative score of 0.84 (below the unit threshold). Our model instead correctly predicts the binding with a margin of 0.999.

### Structure-based energy model performance on microarray data

A structure-based energy model has been developed by Wunderlich *et al.*, which predicts the interaction energy between the SH2 domains and the peptides [230]. This energy model was trained on a large scale microarray experiment [165] (our dataset II), and a significant energy

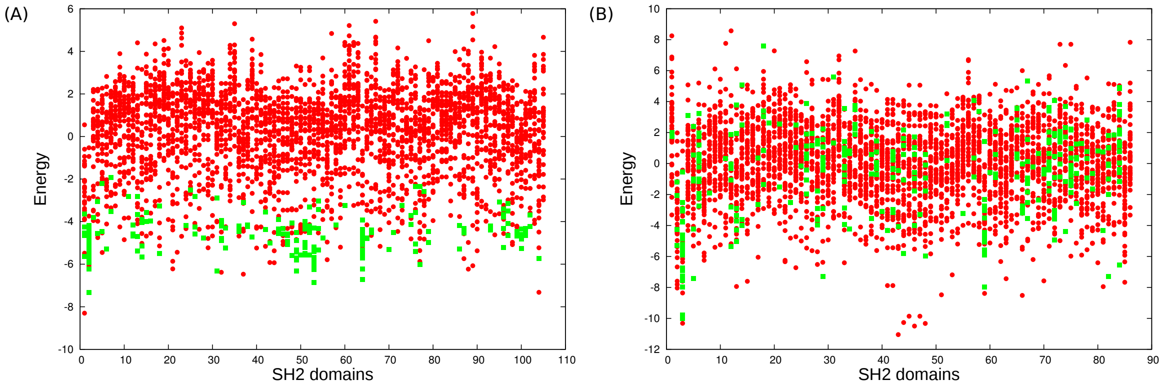


**Figure 4.4.:** Binding and non-binding energy comparison with different microarray data. (A) Plots for the binding and non-binding energies derived from dataset II, indicates there are clear difference between the binding (red dots) and non-binding interactions (green boxes). (B) With the data derived from dataset III, surprisingly, we observed that there is no clear differences between the binding (red dots) and non-binding (green boxes) interactions. The energy calculation program was kindly provided by Zeba Wunderlich [230]. This figure is taken from [P4].

difference between binding and non-binding interactions was observed [230]. When we apply this energy model on the dataset II, not surprisingly, we obtain the results reported in [230]; namely TPR 0.90 and FPR of 0.06. More precisely, we could determine the threshold value that achieves the reported classification results. In Figure 4.4.A, there is a clear energy difference between the binding and the non-binding pairs. The software was kindly made available to us by Zeba Wunderlich.

However, when we apply the same energy model (trained on dataset II) on dataset III [44], we obtain quite a different result. Figure 4.4.B clearly indicates that there are no prominent energy differences between the binding and non-binding pairs. Moreover, we observed in this case, there is no threshold that can significantly discriminate between the binding and the non-binding cases (see also AUC ROC results in A.1.4). This seems to indicate an over-training issue with a consequent inability of generalization to a different experimental setup.

### 4.2.7. Discussion

SH2-peptide interactions are an important component of cell signaling. Because of the limited availability of experimentally proven interactions, machine learning approaches have to be used in order to generalize to combinations that have not been experimentally investigated. High-throughput experimental methods seem to be a perfect data source for training these models. There are, however, two kinds of problems in these data: (i) a significant noise component and (ii) quite an imbalance between confirmed interactions (positive data) and experimentally proven non-interactions (negative data). In addition, current state-of-the-art models for SH2-peptide interaction prediction are based on linear models, which are not capable of handling complex interactions patterns.

In our work, we propose a model that tackles these issues. On the one hand, we propose an iterative re-balancing strategy to compensate the imbalance problem. On the other hand, to model complex interaction patterns, we use a polynomial kernel support vector machine, and we avoid overfitting issues employing a regularization scheme.

We used three high-throughput data: two derived from microarray experiments and one from a peptide array library experiment. We carefully compared our approach with state-of-the-art tools, namely `SMALI` and an energy based structural model, achieving a significantly better generalization performance (measured as cross-validated AUC ROC and AUC PR). This result was additionally confirmed on a manually curated database (`PhosphoELM`) of experimentally validated SH2-peptide interactions. Finally, we performed a genome-wide prediction of human SH2-peptide interactions and report some novel interactions between SH2 domains and tyrosine-phosphorylated proteins (see Section 4.5.2 for details).

## 4.3. SH3-peptide interactions

### 4.3.1. Introduction

For predicting SH3 mediated interactions, we presented a graph kernel-based machine learning approach (see Section 3.3.2 for details). This graph-kernel technique, differently from the PWMs, does not require an initial peptide multiple alignment. Furthermore, by virtue of its non-linearity assumptions, it can adequately capture all types of peptide classes. We build specialized models for the 70 human SH3 domains and achieve much better predictive performance compared to the state-of-the-art method [192]. We show that better models can be obtained when we use information on the non-interacting peptides (negative examples), which is currently not used by the state-of-the art approaches based on PWMs. Moreover, we show how we can leverage the information contained in related domains by building a single comprehensive model for a set of 6 SH3 domains (see Section 3.3.4), which further improve the predictive performance. Although high-throughput datasets are available to train statistical based learning approaches, we note that the presence of spurious interactions in the experimental data (either false negatives or false positives) can severely affect the quality of the induced model. However, in Section 3.3.3, we proposed two methods to generate high-confidence negative interaction data. These negative class instances were then used to train a model in a setting with reduced noise to signal ratio.

### 4.3.2. Dataset compilation

#### Dataset I (high density peptide array data)

In our study, we use the large scale human SH3-peptide interaction data from a high density peptide array experiment [58]. A total of 9192 peptides of length 15 were used in the CHIP experiment. The SH3-peptide interactions that gave a positive signal in peptide chip experiment have been stored in the newly developed interaction database `PepspotDB` [58].

**Table 4.4.:** Summary of the whole data for 70 human SH3 domains. Data available from the high density peptide array experiment of [58]. In brackets, the interactions evidence available in MINT [196]. The table is taken from [P3].

| | # Positive | # Negative | # Unknown |
|---|---|---|---|
| **Peptides** | 2802 | 9188 | 9188 |
| **Interactions** | 16032 (478) | 262883 | 627177 |

From PepspotDB, we have retrieved 16032 non-redundant interactions for 70 human SH3 domains and 2802 peptides. Among them, a total of 478 interactions were also supported by the literature as reported by the MINT database [196] (see Table 4.4).

### 4.3.3. Dealing with false negatives

Traditional methods for peptide characterization rely on generative approaches where the probability of the model (often represented as a motif) is estimated from positive data alone. A typical approach is represented by position weight matrices (PWMs) [192] where the multinomial probability distribution for each position in the sequence is estimated independently via frequency counts. In the machine learning community, it is known that discriminative models have an advantage over generative ones, since they can rely on both positive and negative data; this allows them to better identify the decision boundary for the relevant region of the data-space. While generative methods often require less training examples, they do not achieve quite the same performance [197]. However, when negative data is assumed to be severely affected by noise, or even when the negative data is overly represented, one-class models can exhibit an advantage over discriminative ones. A typical scenario is when dealing with high-throughput experimental results, such as phage display [261], SPOT synthesis [262], or peptide array screening [130]. Here, in order to increase the confidence on the measurements, the experimental protocol makes use of stringent thresholds (e.g., requiring agreement on several replicated experiments). In these cases, a large part of what would be labeled as a lack of interaction (negative example), is in fact just a weaker true interaction (positive example). To deal with these cases, we developed two approaches. The first one is a generative approach that makes use of multiple PWMs to model each peptide class. We then select a subset of instances that are not recognized by any specialized PWMs and use those as reliable negative instances to train a binary classifier. The second approach is based on a combination of a one-class and a semi-supervised method.

### 4.3.4. Single domain model with filtered negatives

As detailed in Section 3.3.3, we induced PWMs to model several known classes of binding peptides for each SH3 domain. We used these models to select and filter away all peptides that were experimentally identified as non-interacting but that are recognized by the PWMs as belonging to one of the known classes of binding peptides. In this way, we obtain a total of 262883 confident negative interactions for all 70 SH3 domains (the full list of positive and negative interaction data along with the class balance is given in Table A.2.2). We encode the peptide sequences as described in Section 3.3.1, and induce a support vector machine (SVM) to model each SH3 domain based on the graph kernel. Note that even if here we use a linear SVM, we are in fact inducing a non-linear model w.r.t. the sequence of amino acid residues, i.e., the linear model is aware of higher-order features that capture the dependency between pairs, triplets, etc, of amino acids.

We used a 10-fold *stratified* cross-validation (see Section 3.6.1) in order to evaluate the predictive performance of each model. The hyper-parameters of the method were optimized in each fold by using a 5 fold cross-validation over the training set. Specifically, we optimized the radius parameter $r \in \{1, \ldots, 8\}$ and the distance parameter $d \in \{1, \ldots, 8\}$ for the graph kernel. The SVM model is induced using the Stochastic Gradient Descent (SGD) approach championed by [263]. The optimal values are achieved at $r = 6$ and $d = 8$ for most of the domains. We report the performance measures in Table A.2.3, and on an average, we obtain a remarkable 0.73 AUC PR and 0.94 AUC ROC.

As for run times, since the NSPDK has essentially a linear complexity when dealing with bounded degree graphs, we report the estimated average time per instance: 0.07 sec/instance on an ordinary 2.33GHz Intel Core2 Duo CPU. This time includes the file upload in main memory, the graph feature generation, and the parameters fitting of the model via the SGD. In practice, this means that we can generate a model given 1K peptides in one minute, or equivalently, a model for a proteome-scale 100K peptides dataset in less than two hours on a desktop machine.

We note that at times, we suffer from the high imbalance problem. For certain domains (e.g., CSK, DLG1, FISH, GRAP2-1, RUSC1, STAM2 etc.), the ratio between the available information for positive interactions and negative interactions is above 100. It is known in the machine learning literature that severely imbalanced class distribution negatively affects the performance of adaptive predictors [210], since the tuning algorithms are generally biased towards the majority class. In our case, the majority class is the negative class, which implies a low sensitivity (true positive rate).

### Comparison with state-of-the-art PWM approach

We have compared our results with a recently developed tool based on PWMs called Multiple Specificity Identifier (MUSI) [192]. Even if the tool tries to increase the modeling complexity by replacing a single PWM with multiple PWMs, it remains in essence a linear model, and therefore still suffers from the issues detailed in the Chapter 2, namely the inability to model dependency between features and the fact that it requires an initial error-prone peptide alignment phase. We have used exactly the same experimental setup as in our approach. In Figure 4.5, we report the comparative results w.r.t. AUC PR and AUC ROC performance measures for all 70 human SH3 domains. On average, MUSI achieves a non-competitive 0.27 AUC PR and 0.69 AUC ROC.

We were curious to see how our method performs on the same experimental dataset used by Kim *et al.*, and hence we collected the interaction data reported in the paper [192]. A total 2457 unique positive interactions were available for the SH3 domain from SRC protein. Since the interaction peptides were identified by the phage display experiment, we could only get the positive interaction data. For preparing the negative interaction data, we have taken three different strategies: (i) we considered the filtered negative data used in our study, (ii) we prepared random negatives automatically generated by $rand()$ function in `Perl`, and (iii) we prepared the random negatives generated by the same strategy as described above with PxxP core, since SH3 domain binding peptides normally contain this core motif (see Section 1.4.2 for binding specificity of SH3 domains). Finally, we have performed a *stratified* 10-fold cross-validation using same parameter ranges ($r \in \{1, \ldots, 8\}$ and $d \in \{1, \ldots, 8\}$) for optimization and report AUC PR and AUC ROC performance measures for all these three datasets. In this experiment, our approach achieved a much higher performance than `MUSI` tool. This would add another layer of confidence to the performance of our models. We also compare the performances of our graph kernel approach and `MUSI` on our original dataset along with these three datasets (see Figure A.2.1).

Note that the problem of generating the initial alignment was also tackled in a recent publication by [258]. They identify multiple specificities in peptide data by performing two essential tasks simultaneously: alignment and clustering, and therefore find biologically relevant binding motifs that cannot be described well with a single PWM. Our approach sidesteps these issues altogether, as we just make a model based on all available peptide features (achieving, at the same time, a speed up of several orders of magnitude in run times).

### 4.3.5. Single domain model with unfiltered negatives

Training and testing systems using only high-confidence negative interactions can, in principle, induce a bias that alters the comparison between methods. To rule out such a case, we perform an additional experiment where we do not filter in any way the negative data. We employ the same setup as in previous experiments (i.e., *stratified* 10-fold cross-validation) using the same parameter ranges ($r \in \{1, \ldots, 8\}$ and $d \in \{1, \ldots, 8\}$) for optimization. In Figure 4.5, we report the comparative results w.r.t. AUC PR and AUC ROC performance measures for all 70 human SH3 domains. The graph kernel approach achieves an average AUC PR 0.35 and 0.90 AUC ROC. Under the same conditions, `MUSI` achieves a non-competitive AUC PR 0.04 and AUC ROC 0.58. This result confirms the advantages of the proposed discriminative graph-based method. Note that the large difference in the performance w.r.t. the filtered case is due to (i) the imbalanced class distribution (some are more than 1:100) and (ii) the presence of a possibly large portion of false negatives.

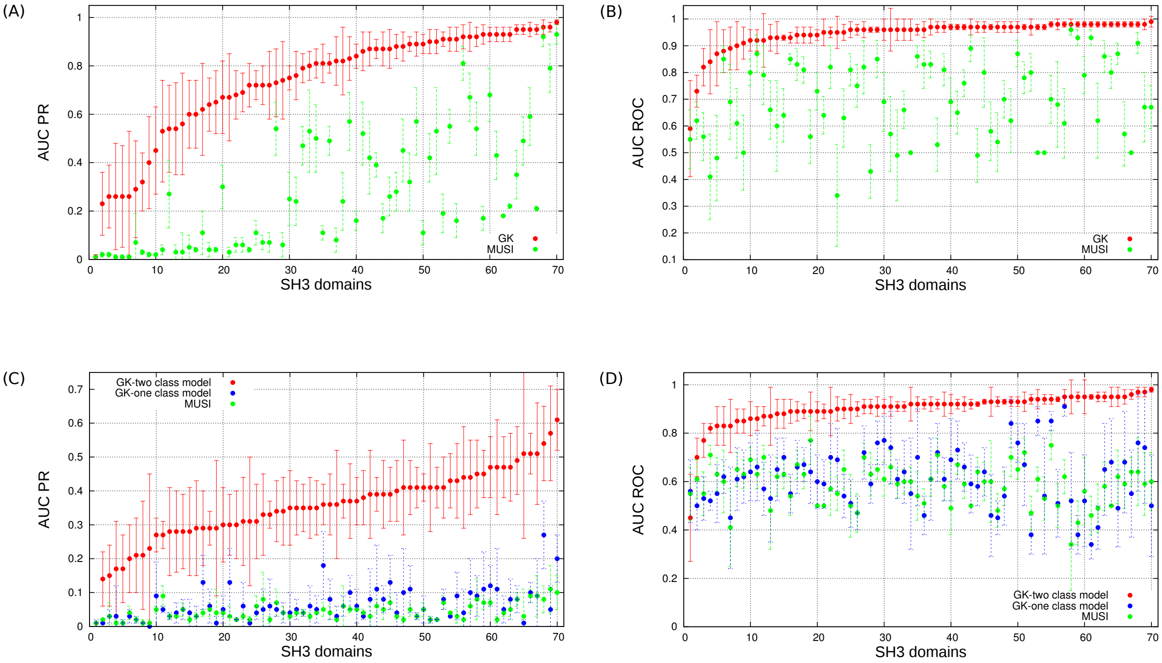


**Figure 4.5.:** A 10-fold cross-validation performance. (A + B) Comparison when using filtered negative interactions for Graph Kernel (GK) and MUSI. (C + D) Comparison with non-filtered negative interactions for binary class Graph Kernel (GK), one-class Graph Kernel, and MUSI. The error bars represent respective standard deviation. The domains are sorted by increasing average performance for the Graph Kernel method. The figure is taken from [P3].

### 4.3.6. Single domain one-class model with semi-supervised filtered negatives

In order to test how important the precise information on true negatives (i.e., peptides that do not interact with the domain) is, we employed the one-class and semi-supervised technique described in Section 3.3.3. The key idea here is to make use of information based primarily on the positive interactions to characterize the binding peptides; instances that are not well recognized by the model are then assumed to be negative. Once again we operate in the same setup as for the unfiltered negatives experiment. In Figure 4.5, we report the comparative results w.r.t. AUC PR and AUC ROC performance measures for all 70 human SH3 domains. The one-class approach achieves an average AUC PR 0.063 and 0.61 AUC ROC. Although this result is statistically significant (according to a Wilcoxon Matched-Pairs Signed-Ranks Test, with p-value = 0.0003), the magnitude of the result lets us conclude that using a generative approach to model protein-peptide interactions is non-competitive w.r.t. discriminative approaches.

### 4.3.7. Multi-domain model and evaluation

As detailed in *Multiple Domains Modeling* in Section 3.3.4, we aligned six domains (SH3 domains for FYN, BTK, HCK, FGR, SRC, and LYN proteins) with the MUSCLE tool [42]. We used the SVM$^{light}$ [232] software to train a Gaussian SVM over the explicit sparse feature

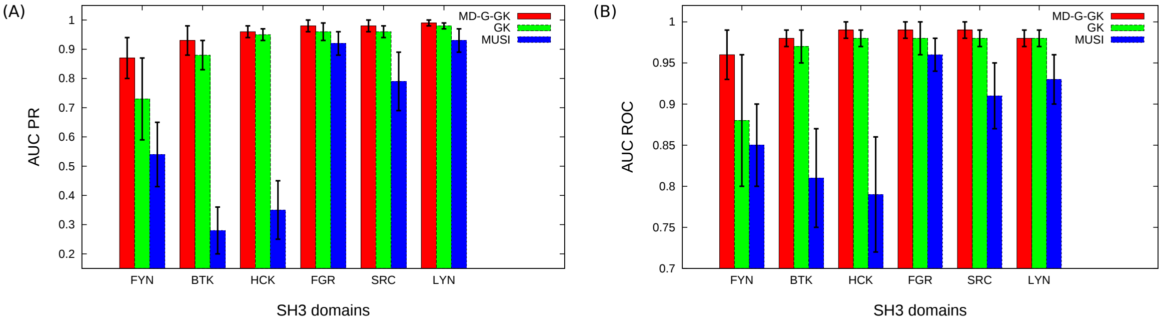


**Figure 4.6.:** (A) Precision-recall curves and (B) AUC ROC curves for the Multi-Domain Gaussian Graph Kernel (MD-G-GK), the Single Domain Graph Kernel (GK), and the MUSI tool for 6 related SH3 domains. The error bars represent respective standard deviation. The figure is adopted from [P3].

encoding of peptide and domain sequence pairs. We evaluated the predictive performance using a 10-fold cross-validation over the six domain sets using the filtered negatives as specified in Section 3.3.3. The value for the Gaussian width was optimized on an internal 20% validation set over the range $\gamma \in \{.001, .01, .1, 1\}$ and the trade-off parameter $C \in \{1, 10, 100\}$, while the values of $r$ and $d$ for the graph kernel were fixed at the optimal value obtained in the previous experiments of $r = 6$ and $d = 8$. In Figure 4.6, we report the AUC PR and the AUC ROC, achieved by single domain model, multi-domain model, and MUSI, for each of six SH3 domain. It is clearly observed that the multi-domain model performs better than the single domain models. As a baseline, we trained (and evaluated in an analogous setting) the six models independently on each domain, both using a linear kernel and a Gaussian compounded kernel. In Figure A.2.2, we report the sensitivity and the specificity, respectively. The experimental result confirms our intuitions: sharing information across related domains increases the predictive performance, mainly due to an increase in sensitivity. We also note that the difference between models trained over single domains when using the linear kernel or the Gaussian one is statistically not significant. This result is also in line with our expectations, since the dependency between features is fully captured by the pairwise neighborhood subgraph features, leaving no margin of improvement to the non-linearity implemented by the *kernel trick*. With radius $r = 6$ and distance $d = 8$, in fact, the kernel generates features spanning the whole sequence.

Finally, we report the performance of the joint model when trained over the six domains, but tested over a novel albeit related LCK dataset. In this experiment, we are asking to predict the specificity for a novel domain given only the information about the alignment of this domain to the overall consensus alignment. The model achieves an average AUC PR 0.85 and AUC ROC 0.96, with a very high sensitivity 0.91 and specificity 0.96. The interesting finding is that the results are better than those obtained by training a model on the LCK protein alone; in this case, in fact, we obtain an average AUC PR 0.86 and AUC ROC 0.94, with a quite low sensitivity 0.55 and a high specificity 0.99. To understand the result, note that in the case of the LCK domain, we have experimental evidence only for 150 positive interactions, while the dataset for the six domains has a total of 910 non-redundant peptides involved in positive interactions. Thus, the experimental results support the hypothesis that, at least in the LCK case, the domain alignment is sufficient to characterize the peptides binding model and to achieve, therefore, a higher overall sensitivity.

### 4.3.8. Comparison with other predictive methods

We have shown our method performs better than the state-of-the-art approach. However, we thought that there are a few questions which still need to be answered. Thus, we performed additional experiments to answer two questions: (i) how well does a classical string kernel fair on this task? and (ii) how important is the alignment phase?

To answer the first question, we have used a $k$-mer string kernel on the amino acid sequence information only without any "abstract information" (like charge or hydrophobicity) in the feature encoding. The $k$-mer kernel simply extracts all substrings up to size $k$ and compares the resulting histograms between two sequences. We identified the optimal $k$ value via an internal 10-fold cross-validation on a validation set (30% of the training set). As for the evaluation, we tested the range of values $k \in \{1, \ldots, 8\}$ and obtained the best performance for $k = 2$. Predictive performance is reported as the average AUC PR and AUC ROC for all 70 SH3 domains (see Figure A.2.4) on a *stratified* 10-fold cross-validation.

To answer the second question, we aligned the peptides with two methods and thus obtained a fixed size vector encoding. Afterwards, we have fair linear and Gaussian predictive models under the SVM loss. For the alignment, we used (a) the `MAFFT` [195] tool allowing "gaps" within the sequences; and as a second strategy we used (b) the `MUSI` approach [192] which enforces "zero gaps" within the peptide sequences (internal gaps are supposed to be biologically not very plausible).

Model hyper-parameters ($\gamma$ for Gaussian kernel and the trade-off parameter $C$ for both Gaussian and linear kernel) have been selected once again by internal 10-fold cross-validation over a 30% validation set, in the range $\gamma \in \{.001, .01, .1, 1\}$ and $C \in \{1, 10, 100\}$, respectively. The average AUC PR and AUC ROC for all 70 SH3 domains over a 10-fold cross-validation are reported in Figure A.2.4.

As a result of the first question, the $k$-mer string kernel is outperformed by our approach, AUC PR = 0.60, AUC ROC = 0.92 vs AUC PR = 0.73 and AUC ROC = 0.94. This confirms the intuition that using physico-chemical properties in the feature definition can adequately model cases that would otherwise be poorly covered by a sufficient number of sequences.

The result of the second question indicates that the Gaussian model with "zero gaps" aligned sequences performed better than other methods, although not as well as our approach. This confirms the intuition that higher-order dependencies are useful for a better modeling of binding specificities. The importance of the "zero gaps" approach is also confirmed. We note, however, that when we compare the simple $k$-mer string kernel and the Gaussian model, they perform similarly, achieving AUC PR 0.60, AUC ROC 0.92 and AUC PR 0.63, AUC ROC 0.89, respectively.

We conclude that in order to achieve top performance, we need to consider systems that are alignment free and that can exploit dependencies between amino acid positions.

### 4.3.9. Discussion

SH3-domain are probably the most widespread class of protein recognition modules, which constitute a very important class of protein-protein interactions, involved in many cellular processes. We presented a computational approach to predict domain-peptide interactions using available high-throughput data. The method is an alignment-free approach based on an efficient graph kernel.

Current methods for protein-peptide interaction often require an initial multiple alignment of the bound peptides. Since this is an error-prone process (especially in the case of SH3-domains, where peptides are proline-rich), one risks to introduce a significant amount of noise and obtain under-performing models (sec Section 2.3.1). In addition, current methods are often linear models (e.g., PWMs) and are, therefore, not able to represent higher-order dependencies between amino acid residues. Non-linear methods exist but have to deal with the high model complexity resulting from exponential number of higher-order dependencies achievable even for relatively short peptide sequences. If one uses the full alphabet of 20 amino acids, it becomes hard to gain sufficient data for a correct estimation of these complex models. One common solution is to use a reduced alphabet where each letter represents an entire amino acid class. This strategy, however, leads to inferior performance, especially when specific amino acids are indeed preferred at specific positions. An alternative approach is to determine important interaction first by using resolved 3D domain-peptide structures. The major obstacle for the wide-spread application of this approach, however, is the limited availability of such structural data.

In our work, we employ a different approach. We consider an alignment-free approach based on a graph representation of the peptide sequence where different abstraction levels are available in a unified way. By applying an efficient graph-kernel method, we were able to model higher-order dependencies that span different abstraction levels (e.g., a feature could represent a specific residue that has to be three positions to the right of a hydrophobic residue). The regularization provided by the SVM optimization scheme finally ensures that the model complexity is appropriately controlled, and that only the features relevant for the task at hand are selected. Discarding the abstraction information, i.e., using only the amino acid code information, leads to a statistically significantly lower sensitivity (see Figure A.2.4). This confirms the intuition that using physico-chemical properties in the feature definition can build better predictive model. It was also important to optimize the encoding order, therefore, we performed an experiment with different encoding order and proposed the best order to represent our graph (see Figure A.2.3).

Although NSPDK graph kernel approach has been previously used for clustering RNA-structures [246], here differently from the RNA or molecular case, we do not have an obvious and natural way to encode the information as a graph. The guiding principle behind the choice of the proposed feature encoding, is to add "abstract information" (like charge or hydrophobicity) in a somewhat "soft" and incremental way. Rather than using an extended alphabet and maintaining a *sequence* encoding, the proposed *graph* encoding allows us to obtain features that are increasingly specialized. We have experimental evidence that a different choice in the ordering of the abstract information would yield suboptimal results, which becomes evident in the presence of imbalanced data (see Figure A.2.3). Additionally, we have investigated the performance of a string kernel (the $k$-mer kernel) along with other types of kernels, applied to the pure amino-acid sequences (i.e., without any additional information). Also in this case, there is an evident drop in the performance (see Section 4.3.8 and Figure A.2.4).

Interestingly, the experimentally cross-validated optimal parameter values ($r = 6$, $d = 8$) suggest that very higher-order amino acid dependencies are indeed required to obtain the best predictive performance and therefore linear models are inadequate. Another common practice is to employ generative models, i.e., models that try to capture the density distribution of the interacting peptides only. We showed that using one-class approaches is suboptimal, even when considering models more expressive than the commonly used linear PWMs. The average predictive performance of a graph kernel based domain specific model that is trained in a discriminative fashion is 0.35 AUC PR compared to 0.06 AUC PR when trained in a one-class way (see Figure 4.5).

We tried to address the problem of selecting high quality negative data. The issue is known in literature [216, 217]. In the application of domain-peptide interaction, it has been shown that the common practice of generating negative instances by randomly shuffling peptide sequences, simply leads to a decreased predicted performance, as these instances do not resemble real biological sequences, and are not therefore useful to determine useful class boundaries [217]. We note, however, that a decreasing performance is proportional to the level of class imbalance. When the ratio of negative instances versus positive ones is within 10 fold, we maintain an AUC PR 0.8, but for ratios greater than 100, performance drops to AUC PR 0.4 and lower (see Figure A.2.5 and A.2.6).

We showed how the flexible graph kernel approach allows the induction of multi-domain models. These models can leverage experimentally verified binding interactions on related domains and achieve high predictive performance even on domains for which no training data was available. Finally, we have performed a genome-wide analysis for uncover novel SH3 mediated interactions (see Section 4.5.2 for details). As for the future work, given the computational efficiency of these models (a single domain model can be trained on 100K sequences in less than two hours), we plan to provide a comprehensive set of predictors for all protein domains for which high-throughput data is available.

## 4.4. PDZ-peptide interactions

### 4.4.1. Introduction

In this section, we present a cluster-based prediction of PDZ-peptide interactions for human (*Homo sapiens*), mouse (*Mus masculas*), fly (*Drosophila melanogaster*), and worm (*Caenorhabditis elegans*) using a machine learning approach. The importance of our method is five fold: (i) clustering of a very large set of PDZ domains based on their sequence identity. This comprehensive study allowed us to construct specialized models for 43 PDZ families, covering 226 PDZ domains, which are more accurate compared to the state-of-the-art and offers models for the largest set of PDZ domains to date; (ii) the data obtained from high-throughput experiments are often found to a lack of non-interacting data (i.e., negative data) and thus lead to a great class imbalance problem. We already know that the performance of many machine learning methods are significantly poorer on highly imbalanced data [235]. To deal with this issue, we employed a semi-supervised machine learning approach to identify high-confidence negative interactions in an analogous fashion, described in Section 3.2.3; (iii) we allowed the dependency between the amino acid positions in the binding ligand; (iv) we built two types of models, one sequence-based and one based on contact information from reference structures, and compared the performance of these models; and (v) finally, we performed a genome-wide analysis for 101 and 102 PDZ domains from human and mouse, respectively, and uncovered novel and biologically meaningful PDZ-peptide interactions.

### 4.4.2. Dataset compilation

For retrieving all the annotated PDZ domains from human, mouse, fly, and worm proteomes, we used `UniProtKB/Swiss-Prot` database, which is a well known manually curated and reviewed database [19]. At the time of analysis, the `UniProtKB/Swiss-Prot` database (release 2013-01) contained 20248 human (*H. sapiens*), 16597 mouse (*M. masculas*), 3182 fly (*D. melanogaster*), and 3382 worm (*C. elegans*) proteins. A large set of 548 PDZ domains, comprising 271 human, 234 mouse, 27 fly, and 16 worm PDZ domains, were derived.

#### Dataset I (microarray data)

We used a protein microarray screening data to analyze the specificity of PDZ domains, comprising 157 mouse PDZ domains and 217 fluorescently labeled genome-encoded peptides [164]. The initial interaction data derived from microarray screening was further analyzed by fluorescence polarization. Apparent equilibrium dissociation constant ($K_D$ value) was applied to determine the positive and negative classes [164]. A total of 731 positive interactions and 1361 negative interactions were derived that involved 85 mouse PDZ domains and 181 peptides using a $K_D$ cutoff of 100 $\mu$M. We used the same $K_D$ cutoff value as mentioned in [164].

### Dataset II (phage display data)

For the human phage display experiment, we considered a total of 1389 interactions that involved 54 human PDZ domains and 1211 peptides [97]. Note that this experiment provides only positive interaction data. Thus, we did not have any negative interaction data for this dataset.

### Dataset III (curated data)

From PDZBase [264], which is a high quality known PDZ-peptide interaction database, we extracted non-redundant 201 interactions, comprising 94 domains and 115 peptides. We considered interaction data only from human, mouse, fly, and worm. Note that PDZBase also contains only positive interaction data and hence no negative interaction data was available in this database.

## Domain-peptide complex structures

For retrieving all the available PDZ-peptide complex structures, we used Protein Data Bank (PDB), which contains experimentally solved protein structures [265]. At the moment of analysis, PDB contained 55 PDZ-protein and/or PDZ-peptide complex structures comprising 47, 5, and 3 structures from human, mouse, and fly, respectively. Note that we were unable to find any PDZ-peptide complex structure for worm. After filtering according to available interaction data, we were left with 21 human, 5 mouse, and 3 fly PDZ-peptide complex structures.

We have combined all the positive and negative interaction data from dataset I, dataset II, and dataset III. Five C-terminal residues of the peptides were considered, since they are the most important for determining PDZ domain specificity [92, 97]. Finally, we retrieved a total of 3592 interactions involving 194 domains and 1437 peptides.

## Tree of PDZ domains

In recent years, enormous amounts of interaction data have been generated by various high-throughput experiments and thus computational methods are invaluable to analyze these data. One of the major problems while analyzing these data is sufficient amounts of data may be available for certain domains, but completely missing or much less available for another domain. For example, there are only two positive interactions for PDZ 1 and PDZ 2 domains of human DLG2 and DLG4 available in the literature. To overcome this limitation, our first goal was to combine the PDZ domains that are similar in substrate specificity and therefore build a single classifier for these similar domains. Hence, this approach enables us to make separate models for each domain family.

First, we aligned all available PDZ domains (human, mouse, fly, and worm) annotated in UniProtKB/Swiss-Prot by using MAFFT and built a phylogenetic tree [195]. We then

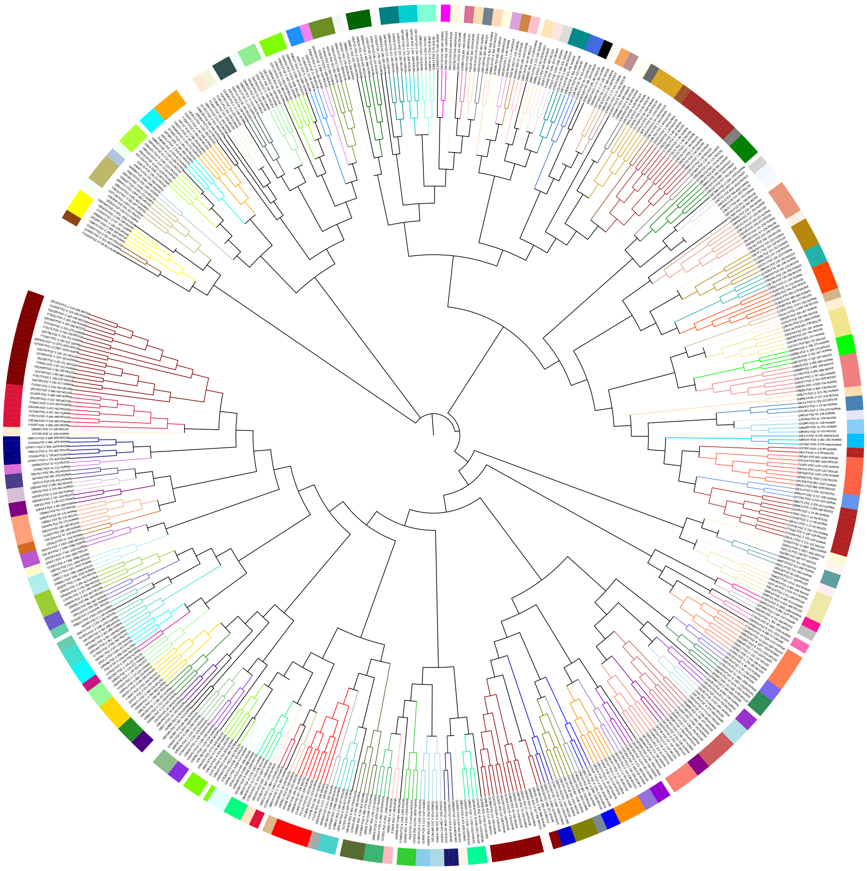


**Figure 4.7.:** Clustering of PDZ domains. Phylogenetic tree of all available PDZ domains from human, mouse, fly, and worm. The `MCL` clustering output was mapped onto the phylogenetic tree. A total number of 138 PDZ families are presented by 138 colors. `iTOL` was used for the visualization [32]. This figure is taken from [P2].

clustered all the similar PDZ domains based on their sequence identity by using Markov clustering algorithm (`MCL`), which is a fast and powerful algorithm for clustering biological sequences [254]. 50% sequence identity was set for the cutoff value, as previous research showed that the PDZ domains with more than 50% sequence identity have similar binding specificity [200] (see Section 3.4.1 for details). All the available PDZ domains (548) were classified into 138 families. Out of all 548 domains, we were unable to classify 33 PDZ domains since the sequences are too diverse. The biggest family consists of 20 PDZ domains from human, mouse, and fly. Finally, we have mapped the 138 families on the phylogenetic tree of all PDZ domains for better visualization (see Figure 4.7). Additionally, we have described the peptide preferences for each PDZ domain family. Amino acid composition of the binding peptides was visualized using sequence logos [179], showing the amino acid enrichment at each position in the binding peptides. See Figure A.3.2 for the ligand binding specificity of each PDZ domain family.

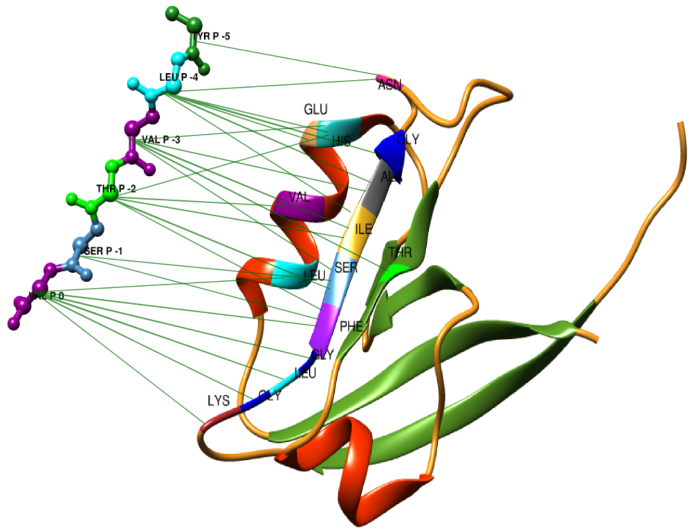


**Figure 4.8.:** PDZ-peptide complex structure. Representative PDZ-peptide complex structure (`PDB`-id: 4G69) for PDZ family 1. 2nd PDZ domain from human DLG1 binds to the C-terminal peptide of human APC protein. Green lines indicate the binding pairs with distance less than 4.5 angstroms. `UCSF Chimera` was used for the visualization [48]. This figure is taken from [P2].

### Modeling

We used two strategies for modeling our data: (i) a purely sequence-based approach and (ii) a contact-based modeling that uses structural information.

### 4.4.3. Sequence-based data modeling

For the sequence-based modeling approach, we followed the literature and considered five C-terminal residues of peptide sequences as an input, where the position of C-terminal residue is given at P0 and going upstream P−1, P−2, and so on. To define the positional features, we extracted amino acids from peptides and mapped them into a binary vector $x$ living in a $20 \times 5 = 100$ dimensional space (see Section 3.4.2 for details). Families with at least 10 positive interaction data were considered for modeling. In summary, we built models for 43 families covering 226 PDZ domains.

### 4.4.4. Contact-based data modeling

The contact-based modeling approach combines the peptide sequence information with PDZ-peptide complex structure information. We followed a similar approach taken by Chen *et al.* in 2008 [88]. However, we did not use only one reference structure for all domains. Instead, we used a specific reference structure for each family by selecting one representative domain-peptide complex structure for each family from the `PDB` database [265]. For these domain-peptide structures, we considered only the position pairs (one amino acid from the domain and another from the peptide) with a distance less than 4.5 angstroms (see Figure 4.8). The important position pairs were separately derived for each PDZ domain family.

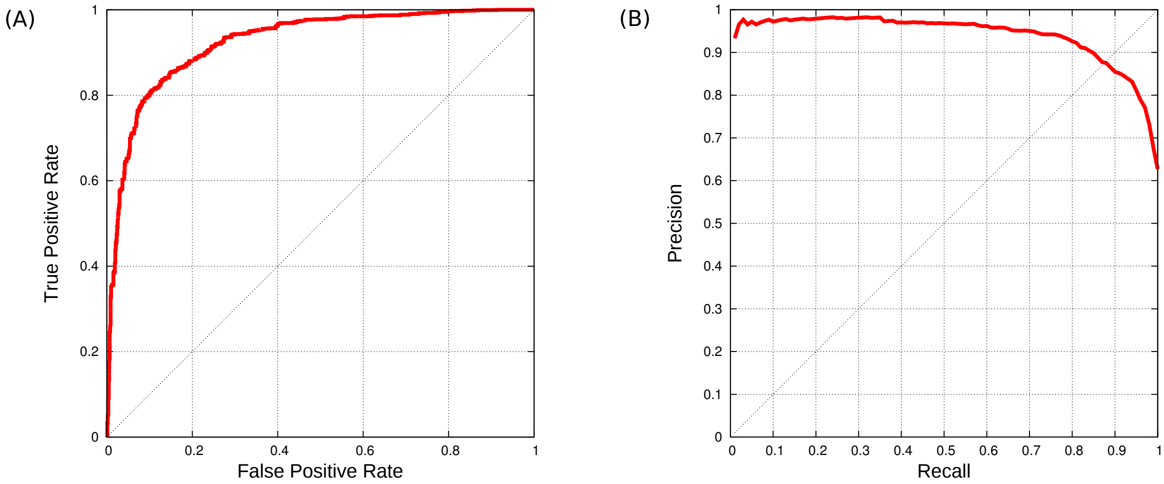


**Figure 4.9.:** (A) The AUC-ROC and (B) the AUC-PR curve obtained by sequence-based feature encoding method. This figure is taken from [P2].

All position pairs were then encoded in a binary vector of size 400 × $n$, where $n$ is the number of binding pairs (see Section 3.4.2 for details). We concatenated sequence-based features with contact-based features and finally, we built models for 10 families covering 70 PDZ domains.

### 4.4.5. Predictive performance evaluation

For the sequence-based approach comprising models for 43 families covering 226 PDZ domains, only 22 families covering 136 PDZ domains met this criteria and therefore were used in cross-validation. The hyper-parameters (i.e., $\gamma$ and the cost parameter $C$) for each fold were optimized using 5-fold grid search method over the training sets. See Table A.3.1 for the performances of all 22 families. We computed area under the ROC curve (AUC ROC) and area under the precision and recall curve (AUC PR) for the 22 families. Using sequence-based feature encoding, we achieved a very good average AUC ROC of 0.92 and AUC PR of 0.94 (see Figure 4.9).

For the contact-based feature encoding method comprising of initially 10 families with 70 PDZ domains, only 6 PDZ families covering 39 PDZ domains met the selection criteria for the cross-validation. Surprisingly, no significant differences were observed when we compared the performances (AUC ROC and AUC PR) of sequence-based and contact-based approaches (see Table A.3.1, A.3.2, and Figure A.3.1). Therefore, we can conclude that the peptide sequence information is sufficient to define the binding specificity of a PDZ domain on the current available data.

### 4.4.6. Benchmarking of existing methods

We compared our results with two state-of-the-art tools, namely `MDSM` (multi-domain selectivity model) [164] and `DomPep` [200], on an independent test set. The independent test set

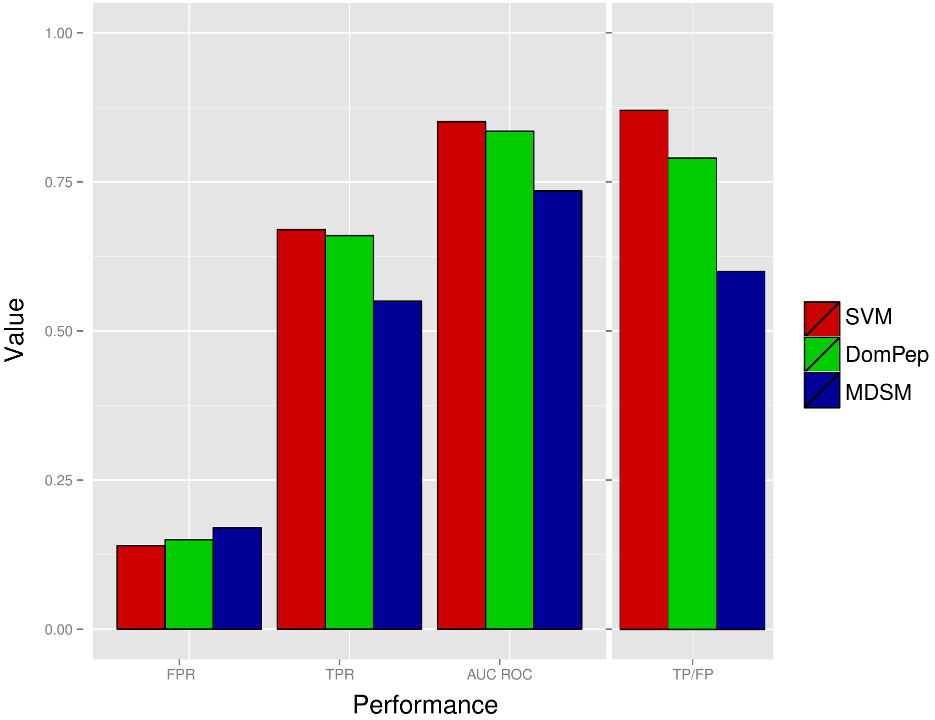


**Figure 4.10.:** Performance comparison of three different tools on an independent test set. Red, green, and blue bars indicate the predicted performances by our tool (SVM), DomPep, and MDSM, respectively. The figure clearly shows that our tool (SVM) achieved better performance. This figure is taken from [P2].

contained 493 positive interactions and 3059 negative interactions that involved 74 mouse PDZ domains and 48 peptides [164]. Among them, we used interactions for 50 PDZ domains that were common in all three methods (MDSM, DomPep, and our method). We make sure that the peptides were not included in our training sets. Our models achieved a true positive rate (TPR) of 0.67, false positive rate (FPR) of 0.14, and AUC ROC of 0.85 with a true-positive/false-positive (TP/FP) ratio of 0.87 outperforming the other two approaches: MDSM achieved TPR of 0.55, FPR of 0.17, and AUC ROC of 0.74 with TP/FP ratio of 0.55; the DomPep achieved TPR of 0.66, FPR of 0.15, and AUC ROC of 0.84 with TP/FP ratio of 0.79 (see Figure 4.10).

In an another experiment, we tested our method with MDSM on a validated dataset. In this case, we could not test DomPep since many of the test instances were present in the DomPep training set, and hence a fair comparison was not possible. The test data was retrieved from an experimentally validated database, called PDZBase [264]. We compared 20 mouse PDZ-peptide interactions derived from PDZBase that were neither included in MDSM nor in our training set. Out of these 20 interactions, we successfully predicted 14 interactions with a true positive rate (TPR) of 0.70, compared to only 4 interactions predicted by MDSM with a true positive rate (TPR) of 0.20. For calculating MDSM score, a unit threshold was defined as the ratio of the original prediction score ($\phi$) over a scoring threshold ($\tau$), specific for each domain [164]. A peptide was then predicted to bind to a PDZ domain $i$, if $\phi_i/\tau_i > 1$.

**Table 4.5.:** SVM and MDSM scores for experimentally validated interactions derived from PDZBase [264]. A peptide is predicted to bind to a PDZ domain, if the score is more than 0 for SVM and more than 1 for MDSM. Bold numbers indicate true positive interactions. This table is taken from [P2].

| PDZ domain | Peptide | SVM score | MDSM score | Pubmed Ref. |
|---|---|---|---|---|
| Cipp-(3/10) | IESDV | **0.44** | -0.7 | 9647694 |
| Cipp-(3/10) | LESEV | **0.30** | -0.62 | 9647694 |
| Cipp-(3/10) | QQSNV | **0.29** | -0.78 | 9647694 |
| Cipp-(3/10) | KEYYV | **0.51** | -0.34 | 9647694 |
| Dvl1-(1/1) | SETSV | -1.27 | -0.74 | 12490194 |
| Pdlim5-(1/1) | DITSL | -0.24 | -0.15 | 10359609 |
| Erbin-(1/1) | LDVPV | **0.99** | 0.61 | 10878805 |
| Magi-2-(5/6) | KESSL | **1.76** | 0.19 | 10681527 |
| MUPP1-(10/13) | IATLV | **1.00** | 0.46 | 11000240 |
| MUPP1-(10/13) | GKDYV | **1.00** | **1.68** | 11689568 |
| NHERF-1-(1/2) | FDTPL | **1.06** | 0.01 | 10980202 |
| LIN-7A-(1/1) | IESDV | **0.33** | 0.29 | 10341223 |
| Lin7c-(1/1) | IESDV | **0.33** | 1.00 | 10341223 |
| ZO-3-(1/3) | GKDYV | **0.99** | 0.09 | 10601346 |
| a1-syntrophin-(1/1) | VLSSV | -1.47 | 0.16 | 11571312 |
| PSD95-(1/3) | LQTEV | **0.38** | **1.41** | 11937501 |
| PSD95-(1/3) | NETVV | -1.35 | **1.19** | 12067714 |
| PSD95-(1/3) | GETAV | -1.32 | **1.23** | 12067714 |
| PSD95-(1/3) | EESSV | -2.23 | 0.77 | 11134026 |
| PSD95-(1/3) | RTTPV | **1.00** | 0.61 | 12359873 |

Table 4.5 lists the scores for all 20 validated interactions as calculated by MDSM and by our method.

The advantages of our approach compared to the aforementioned tools are threefold: (i) using an accurate clustering approach allows our method to achieve a higher domain coverage, (ii) we have employed a powerful semi-supervised learning technique to identify high-confidence negatives, which increases the model quality, and (iii) our approach is based on a non-linear model to address the issue of the dependency between amino acid positions.

### 4.4.7. Discussion

In our comprehensive study, we propose a cluster-based computational method to accurately predict the binding partners of PDZ domains using the support vector machine (SVM). First, we used an efficient MCL algorithm to cluster all PDZ domains from different organisms and thereafter, built prediction models for each PDZ family found by our clustering technique. Our method offers the largest number of prediction models for PDZ domains to date. We showed that our clustering method maximizes the size of training datasets, which is important to build powerful prediction models. In the clustering method, we combined all the PDZ domains that share high sequence identity and therefore have similar binding specificity. There are, however, additional cases where the binding preference is very similar despite a low sequence identity. For example, MAGI1-5 and MAGI3-4 domains share similar specificity despite of low sequence identity (24% in mouse) [200]. Since these cases are hard to detect automatically, we used a conservative approach by considering a threshold of 50% sequence identity. Even using this conservative threshold, we were able to achieve a very good prediction accuracy. We also applied semi supervised learning (SSL) strategy for selecting high-confidence negative data to re-balance our training sets. In our study, we have developed models based on two feature encoding methods: (i) sequence-based and (ii) contact-based methods. Since the sequence-based approach does not depend on domain-peptide complex structures, it covers more PDZ domains, but may fail to predict the binding peptides of a mutated domain with completely different specificity. For mutated domains, we can efficiently use our contact-based approach, which considers binding pairs of domain and peptide, and thus should be able to more precisely evaluate the effect of mutations. Our method is also able to predict the binding peptides of newly characterized PDZ domains. We compared our tool with published state-of-the-art methods and achieved better performance. Finally, we performed a genome-wide analysis to predict several novel interactions for human and mouse PDZ domains (see Section 4.5.2 for details).

## 4.5. Genome-wide predictions and biological insights

### 4.5.1. Introduction

Genome-wide prediction is often used to uncover unknown protein-protein interactions. We have performed a genome-wide analysis for modular protein domains that have been used in our study. Our aim was to identify the novel interactions that have important biological roles. In this section, we describe the setup for genome-wide analysis of three different modular protein domains (e.g., SH2, SH3, and PDZ) and the required prediction filters, which are important to avoid any unlikely interactions. Furthermore, we discuss the possible biological insights of the top interactions predicted by each domain.

### 4.5.2. Genome-wide prediction setup

#### Genome-wide analysis of human SH2 domains

In this setup, we have made use of prior domain knowledge to remove peptides that are not likely to interact. Specifically, we have considered three criteria for the eligibility of a given pair domain-peptide: (i) presence of the tyrosine (Tyr) residue in the peptide, (ii) experimentally verified phosphorylation of the tyrosine in the peptide, and (iii) co-cellular localization of the mature protein that contains the peptide and the protein that expresses the domain.

We have extracted a set of peptides from the `UniProtKB/Swiss-Prot` database [19], which is a well known manually curated and reviewed database. At the moment of analysis, the `UniProtKB/Swiss-Prot` database, release 2012-06, contained 20,225 human proteins with ∼300,000 (298,637) tyrosine containing peptides. The first filter will check whether there is any Tyr residue present in the peptide. The second filter has been implemented using the annotated information from the PhosphoSitePlus database [266]; in this way, we have selected only those phosphotyrosine peptides whose phosphorylation has been experimentally verified. At the moment of analysis, the PhosphoSitePlus database contained 30,228 phosphorylation sites from 10,688 human proteins. We have ignored those peptides that were not present in the UniProtKB/Swiss-Prot database, and finally obtained a total 27,481 phosphorylation peptides out of 9,621 proteins. The third filter was implemented considering the terms relative to the sub-cellular localization hierarchy in the controlled vocabulary of the Gene Ontology (GO) database [267]. In case of multiple cellular locations (e.g., GRB2 protein can be found in nucleus, cytoplasm, endosome, and golgi apparatus [268]), we consider a peptide viable for interaction, if it shares at least one of the terms with the domain. Finally, we ignored proteins (such as SHD/E105251) for which no localization annotation is available.

### Genome-wide analysis for SH3 domains

In order to perform a genome-wide analysis of SH3 domains, we used the manually curated UniProtKB/Swiss-Prot database [19]. We retrieved 20,225 human proteins from UniProtKB/Swiss-Prot database, release 2012-06. For retrieving the peptide sequences, we scan all the available proteins with a window size of 15 and step size of 5. In this way, we have extracted a total number of ∼2M (2,209,474) peptide sequences.

In this analysis, we implemented two different filters: (i) proline-rich and (ii) co-cellular localization. Since SH3 domains are known to interact with proline-rich peptides, we implemented our first filter, namely proline-rich filter, that allows to use several regular expressions to select proline-rich peptides. As a second filter, we implemented co-cellular localization filter in an analogous fashion described earlier in this section to avoid unlikely interactions.

### Genome-wide prediction of PDZ domains

Here, we performed a genome-wide prediction of PDZ domains. In order to do so, we extracted a set of peptides from the manually curated UniProtKB/Swiss-Prot database, release 2013-01 [19]. We retrieved 20,248 and 16,597 proteins from human and mouse proteomes, respectively. The last 5 C-terminal residues were taken from each protein to build the peptide sets separately for human and mouse. In this analysis, we have used prior knowledge to avoid peptides that are not likely to interact with their respective PDZ domains. Therefore, we considered two filters for selecting the probable binding peptides for a given PDZ domain: (i) structural location of the peptides and (ii) co-cellular localization of domain and peptide containing proteins.

Previous study showed that PDZ domains have a tendency to interact with intrinsically unstructured proteins (IUPs) [142], and thus we considered only those peptides that reside in a disordered segment of a protein. For determining the structural disorder of a protein region, we ran the `IUPred` algorithm over the full-length protein sequence to get a disorder score between 0 and 1 for each residue of a protein [269]. Finally, an average score for the last 5 residues (peptide sequence in our study) was obtained to determine putative candidate regions for interaction. A cutoff value of 0.4 was chosen based on the analysis done by Akiva *et al.* [142]. To this end, we ignored all the peptides having the `IUPred` score less than 0.4. As a second filter, co-cellular localization was applied to avoid unlikely interactions in an analogous fashion described earlier in this section.

### 4.5.3. Functional annotation of predicted proteins

For SH2 domains, all eligible domain-peptide pairs were scored by the trained models and ranked according to the SVM scores. Considering the top ranked and most reliable 50 predictions (see Section 4.5.4), we offer the following hypothesis:

(a) The SH2-domain of ABL1 is predicted to bind to Y307 of the adaptor protein GAB1. ABL1 is part of the oncogenic protein BCR-ABL, which is generated by a (9;22) translocation resulting in the so-called Philadelphia chromosome and is found in CML (chronic myelogenous leukemia) [270]. BCR-ABL has been shown to be dependent on GAB adaptor proteins, in particular GAB2. It has been demonstrated that GAB2 in CML cells confers resistance to multiple BCR-ABL inhibitors [271]. The known interaction between BCR-ABL on one side and GAB adaptor proteins on the other side can be described as following: the small adaptor protein GRB2 binds to phosphorylated Y177 on BCR-ABL via its central SH2-domain, and via its SH3-domains it interacts with proline-rich sequences within both GAB proteins, GAB1 and GAB2 [272]. Our finding would suggest a second, so far, unknown mode of BCR-ABL/GAB1 interaction that is GRB2-independent and based on a direct interaction between the BCR-ABL (ABL1) SH2-domain and tyrosine-phosphorylated GAB1.

(b) Our model indicates that the SH2-domain of the adaptor protein CRKL interacts with phosphorylated Y215 of ABL1. Interestingly, CRKL has been found to be one of the predominant substrates of the oncogenic kinase BCR-ABL [273]. This suggests that CRKL is not only a substrate, but also an interaction partner of BCR-ABL. Most likely, the interaction promotes phosphorylation.

(c) TEC-family kinases are multi-domain cytoplasmic tyrosine kinases, which comprise, among others, an N-terminal PH-domain. This PH-domain interacts with the phospholipid phosphatidylinositol-3,4,5-trisphosphate (PIP3), which is generated by PI3K enzymes upon receptor activation [274]. PI3K class IA, which is activated downstream of multiple receptors, such as immune receptors and cytokine receptors, comprises various catalytic and regulatory subunits [275]. Interestingly, our model found that different TEC-family kinases (BTK, ITK, and TEC) via their SH2-domains can interact with various regulators subunits of PI3K class IA: BTK interacts with Y74 of p85$\beta$; ITK interacts with Y464 of p85$\beta$, with

Y467 and Y556 of p85$\alpha$, and with Y199 of p55$\gamma$; TEC interacts with Y74 of p85$\beta$ and Y556 of p85$\alpha$. Since the regulatory subunits of PI3K are necessary to guide the catalytic PI3K subunits to their substrate in the plasma membrane, interaction of TEC kinases with the regulatory subunits would enable them to be close to the newly generated PIP3, which then is necessary for their activation. Using such a mechanism, TEC kinases always could be close to newly generated PIP3 enabling immediate activation.

(d) The inositol-5-phosphatase SHIP1 has been shown to interact with TEC via TEC SH3-domain binding to a proline-rich sequence in the C-terminus of SHIP1 [276]. Our model suggests that there is a second mode of interaction between SHIP1 and TEC, namely between the SH2-domain of TEC and the phosphorylated Y221 of SHIP1. Such a mode of interaction would be called *"bidentate"* and has already been found for the interaction between SHIP1 and one of its main interaction partners, the adaptor protein SHC. In that case, the PTB-domain of SHC binds to a phosphorylated tyrosine within the C-terminus of SHIP1 and the SH2-domain of SHIP1 binds to a phosphorylated tyrosine within SHC [277]. Using such a *bidentate* mode would clearly strengthen the interaction between the two partners.

(e) The inositol-5-phosphatase SHIP1 counteracts PI3K signaling via its centrally located catalytic domain, hydrolyzing the phospholipid PIP3 [277]. Moreover, it has been demonstrated to negatively regulate p21Ras signaling via complex formation with the adaptor protein DOK1 and the p21Ras GTPase activating protein RASGAP [278]. So far, such an interaction or function has not been described for the second family member, SHIP2. Interestingly, our model suggests the interaction of the SH2-domain of SHIP2 (INPPL1) with phosphorylated Y650 of another p21Ras GTPase activating protein, RASA2. This would suggest that both SHIP proteins can realize comparable functions, however, using different modules. The qualitative outcome might be the same, although regulation might be differentially accomplished.

(f) Induction and regulation of calcium mobilization downstream of the B-cell antigen receptor is crucial for differentiation and activation of B-lymphocytes. It was shown that the tyrosine-phosphorylated adaptor protein DOK3 interacts with the SH2-domain of the adaptor protein GRB2. Stork *et al.* have demonstrated that this DOK3/GRB2 module negatively influences the assembly of the calcium initiation complex and/or inhibits the enzymatic activity of the tyrosine kinase BTK, which is crucial for calcium mobilization to occur [279]. Our data indicated that the SH2-domain of BTK directly interacts with DOK3 phosphorylated on Y398. Though our analysis was performed in the human system and the study by Stork *et al.* was making use of the chicken DT40 B-cell system, sequence comparison suggests that the same tyrosine (Y398 in human and Y331 in chicken [279] could bind to GRB2 and BTK. This would add another layer of complexity to the regulation of calcium mobilization in B-lymphocytes.

We performed a second type of analysis on the same top 50 predictions in order to uncover novel functionalities using the `DAVID` tool [280]. The tool offers the possibility to perform

**Table 4.6.:** Predicted peptides that can potentially interact with more than 40 SH2 domains. The table information is taken from [P4].

| UniProt-id | Position | Peptide |
|---|---|---|
| P05067 | 755-761 | NGYENPT |
| P61106 | 12-18 | FKYIIIG |
| P09211 | 48-54 | CLYGQLP |
| P25788 | 103-109 | FGYNIPL |
| P29350 | 562-568 | DVYENLH |
| Q05397 | 923-929 | KVYENVT |
| P08865 | 137-143 | ASYVNLP |
| P13533 | 552-558 | KLYDNHL |
| P56945 | 10-16 | ALYDNVA |
| O15530 | 374-380 | GNYDNLL |

a term-centric enrichment analysis, which identifies enriched annotation biological terms associated with the predicted proteins, on more than 40 different annotation categories. The smaller p-values indicate higher enrichment. Analyzing the highly enriched results we found, for example, that CRKL interacts with a group of proteins (UniProt-id: Q13480, P42684, Q9UQM7, Q13555, P00519, P42345, Q13554, and Q13557) that play an important role in ErbB signaling pathway (p-value $3.03 \times 10^{-8}$), as reported in the KEGG pathway database [281]. We note that the SMALI tool misses all these associations (see Section 4.5.4).

Finally, we found that some peptides (see Table 4.6) are predicted to interact a-specifically with more than 40 SH2 domains. In addition, we observed 3-phosphoinositide-dependent protein kinase 1 (UniProt-id: O15530) targeted by the most number (34 domains) of SH2 domains that share the same cellular compartment and functions annotated in Gene Ontology (GO) database.

For SH3 domain, after filtering the eligible peptides, we scored them by the trained models and ranked them according to the SVM scores. Finally, we report the top 50 predictions by each SH3 domain (see Section 4.5.4). Among the predictions, we observed a peptide (CKKLSPPPLPPRASI, position 151-165) from Phosphatidylinositol 4-phosphate 3-kinase C2 domain-containing subunit beta (UniProt-id: O00750) was targeted by many SH3 domains (21 domains) that also share the same cellular compartment as annotated in Gene Ontology (GO) database. There are also evidences of interactions between PIK3C2B with GBR2 and PLC$\gamma$-1 reported in STRING database [282]. In addition, we took 478 real interactions reported in the MINT database [196], discarded them from our training set and could recover 397 interactions (i.e., a recall 0.83).

Furthermore, we performed an analysis on these top 50 predictions for each SH3 domain to uncover the novel interaction functionalities using DAVID tool [280]. The tool allows the possibility to perform a term-centric enrichment analysis on more than 40 different annotation categories. DAVID functional annotation chart, which identifies enriched annotation terms associated with the predicted proteins are reported. The smaller p-values indicate higher enrichment (see Section 4.5.4).

Applying the term-centric analysis we have observed some biologically meaningful interactions. For example: (i) SH3 domains from human P85-$\alpha$ binds to a potential group of proteins (`UniProt`-id: P21854, Q08209, Q07890, O00459, and Q6ZUJ8) that play an important role in B cell receptor signaling pathway; and (ii) among the top predictions by the SH3 domain from human BTK protein, more than 50% proteins take a vital role in alternative splicing.

For PDZ domains, the eligible peptides were scored by the trained models and sorted according to their SVM scores. We have observed C-terminal peptide (IETHV) from Connector enhancer of kinase suppressor of ras 2 protein (Q8WXI2-human, Q80YA9-mouse) was targeted by 40 PDZ domains, representing 16 families, in human and mouse. See Table A.3.3 and Table A.3.4 for top five peptides targeted by most number of human and mouse PDZ domains. See Section 4.5.4 for the top predictions for each human and mouse PDZ domains.

### 4.5.4. Availability

All the top genome-wide prediction data and term-centric analysis are freely available for the scientific community.

#### Genome-wide prediction for SH2 domains

Top genome-wide prediction data and term-centric analysis for SH2 domains can be found under the URL:
`http://www.bioinf.uni-freiburg.de/Software/SH2PepInt/Genome-wide-predictions.tar.gz`

#### Genome-wide prediction for SH3 domains

Top genome-wide prediction data and term-centric analysis for SH3 domains can be found under the URL:
`http://www.bioinf.uni-freiburg.de/Software/SH3PepInt/Genome-Wide-Predictions.tar.gz`

#### Genome-wide prediction for PDZ domains

Top genome-wide prediction data for PDZ domains can be found under the URL:
`http://www.bioinf.uni-freiburg.de/Software/PDZPepInt/Genome-wide-predictions.tar.gz`

### 4.5.5. Discussion

We performed a genome-wide analysis of modular domain-peptide interactions and report some novel interactions: as an example, we find that oncogenic protein BCR-ABL (ABL1) may directly bind (not dependent on GRB2) with pY307 of the adaptor protein GAB1. The specificity or false positive rate (FPR) of interaction predictions has been improved by implementing appropriate filters. Furthermore, as for run times, our methods are efficient, since the time complexity is linear. Thus, the genome-wide interaction predictions of modular domains can be achieved with a higher accuracy and in less time.

Chapter 5

# MoDPepInt: an interactive web server

## 5.1. Overview

In this chapter, we describe MoDPepInt (Modular Domain Peptide Interaction), which is a new and easy-to-use web server for the prediction of binding partners for modular protein domains. Currently, we offer models for SH2, SH3, and PDZ domains via the tools SH2PepInt, SH3PepInt, and PDZPepInt. More specifically, our server offers predictions for 51 SH2 human domains and 69 SH3 human domains via single domain models, and predictions for 226 PDZ domains across several species, via 43 multi-domain models. All models are based on support vector machines with different kernel functions ranging from polynomial, to Gaussian, to advanced graph kernels. In this way, we model non-linear interactions between amino acid residues. Results were validated on manually curated datasets achieving competitive performance against various state-of-the-art approaches. The work presented in this chapter is a part of the [P1].

## 5.2. Introduction

In this thesis, we have used state-of-the-art machine learning approaches to build support vector machine (SVM) models that can accurately predict binding specificity. We have integrated our three different tools: SH2PepInt [P4], SH3PepInt [P3], and PDZPepInt [P2], for three different modular domains, namely SH2, SH3, and PDZ, into a unified web-based system called MoDPepInt [P1]. Currently, we offer single domain models for 51 SH2 human and 69 SH3 human domains, and multi-domain models for 226 PDZ domains across human, mouse, fly, and worm. To assess the quality of our models, we have used manually curated interaction data achieving competitive performance against various state-of-the-art approaches (see Chapter 4 for details).

In summary, the unique features of MoDPepInt include (i) a domain-peptide prediction system for SH2, SH3, and PDZ in a single platform and (ii) the largest number of modeled domains (see Table 5.1).

### Availability

The MoDPepInt server is available under the URL: http://modpepint.informatik.uni-freiburg.de/

## 5.3. Application and functionality

### 5.3.1. Input

All tools have a unified input format. Query sequences (up to a maximum number of 500) can be supplied either in a FASTA format or using UniProt database accession numbers. PDZPepInt offers predictions also for domains that are newly developed and/or not comprised in the original 226 PDZ domains; the unknown query domain should be supplied in FASTA format. Multiple query domain sequences can also be provided.

**Table 5.1.:** Domain coverage of the available tools. The table clearly shows that the MoDPepInt has higher domain coverage than other tools. This table is taken from the supplementary materials of [P1].

| Tools | Domains | | | Total | Pubmed Ref. |
|---|---|---|---|---|---|
| | SH2 | SH3 | PDZ | | |
| Scansite | 14 | 13 | - | 27 | 12824383 [181] |
| SMALI | 76 | - | - | 76 | 18424801 [183] |
| DomPep | 97 | - | 189 | 286 | 22003397 [200] |
| SH3Hunter | - | 16 | - | 16 | 16870929 [206] |
| MoDPepInt | 51 | 69 | 226 | **346** | 24872426 [P1] |

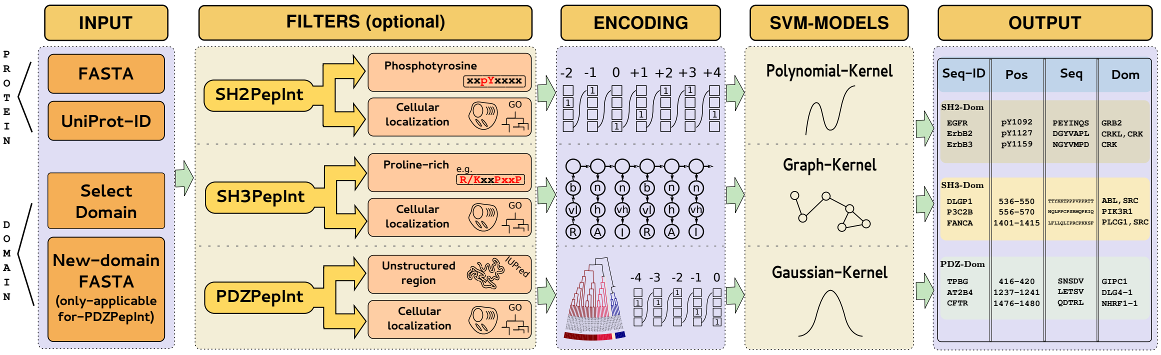


**Figure 5.1.:** Schematic representation of the MoDPepInt pipeline. This figure is taken from [P1].

## 5.3.2. Filters

Several filters are available to increase the predictive accuracy. SH2 domains generally recognize phosphotyrosine (pY) residues of binding proteins. For this reason, in SH2PepInt, we offer a *phosphotyrosine* filter that only considers those peptides whose tyrosine phosphorylation has already been experimentally verified and reported in PhosphoSitePlus database [266].

As SH3 domains mainly bind to proline-rich peptides, in SH3PepInt, we offer a *proline-rich* filter that uses 31 regular expressions to select proline-rich peptides [58].

PDZ domains have the tendency to bind the unstructured C-terminal regions of binding proteins, hence in PDZPepInt, we offer a filter to select for *intrinsically unstructured/disordered regions* based on the IUPred algorithm [269], which selects 5 C-terminal residues with IUPred scores above 0.4 [142].

Finally, a *cellular localization* filter is available for all tools. This filter considers only those interactions where both the protein containing the peptide and the protein containing the modular domain have the same cellular localization according to the Gene Ontology (GO) database [267].

## 5.3.3. Processing and output

An internal queuing system (which currently uses 40 computation nodes) balances the submitted jobs in parallel. MoDPepInt is implemented in C++, Perl, and shell scripting with runtimes typically ranging in the order of a few minutes.

The output for all three tools is formatted as a downloadable table. We report for each domain-ligand protein interaction pair: (i) the sequence ID, (ii) the ligand binding position, (iii) the ligand binding sequence, and (iv) the ligand binding domains. See Figure 5.1 for the schematic representation of the MoDPepInt pipeline.

## 5.4. Meta-web server

In addition to the three specialized servers for SH2, SH3, and PDZ, we implemented the meta-web server MoDPepInt. This meta-web server is to be used in a non-expert mode: (i) no parameters need to be set, (ii) the output comprises predictions for all available domains for SH2, SH3, and PDZ, and (iii) only the five most confident predictions for each domain will be reported. However, the user can easily select one of the dedicated tools for the same input to access the full prediction results and have a finer control over its parametric setting.

## 5.5. Results and discussion

MoDPepInt collects three protein-protein interaction predictive models that can be efficiently tuned using data derived from various high-throughput experimental techniques and thus do not require structural information as in [222, 283, 284]. The resulting models exhibit significant performance improvement in comparison with other existing tools. The main sources of performance improvement are due to: (i) non-linear modeling, where we allow higher-order dependencies between the amino acid positions in the binding peptides, and therefore has an advantage over linear PWM models; (ii) balanced discriminative training where we tackle the class imbalance problem and derive high-confidence negative data; and (iii) dataset pooling where we combine all domains that are similar in substrate specificity. Note that the dataset pooling technique is only implemented in PDZPepInt.

SH2PepInt uses polynomial kernels and it is trained on additional high-confidence negatives obtained via semi-supervised techniques.

SH3PepInt uses graph kernels on a complex representation of both the peptide sequences and of the aligned domains. The adoption of a graph-type representation allows the inclusion of the physico-chemical properties of amino acids, which increases the generalization capacity of the models. Furthermore, the method does not need any prior alignment of the peptides. This is a big advantage since polyproline-rich peptides are hard to align.

PDZPepInt uses Gaussian kernels to train the classifier on the interaction data from highly related domains. In this method, efficient clustering and building multi-domain models help us to leverage the domain-peptide binding information from a limited set of experimental data and extrapolate that information to define specificity for other unseen, but alignable, novel domains.

Once trained, all models can be used to efficiently scan entire proteomes to identify novel interactions with typical runtimes of a few minutes. In addition, we offer a meta-web server to be used in non-expert mode that submits the input simultaneously to all tools and displays a summary of the main results. Overall, MoDPepInt contains largest number of domain-peptide prediction models to date, and allows biologist to investigate the binding specificity of modular protein domains.

# Chapter 6

# Conclusion

Modular protein domains regulate numerous signal transduction pathways by interacting with short linear peptides. In this thesis, we have introduced three different machine learning-based methods for the prediction of modular domain-peptide interactions. The main aim of this work was to build efficient models that can accurately predict the modular domain mediated interactions and can circumvent the limitations of existing methods.

In Chapter 1, we have provided a general introduction of modular protein domains and their relationship to different cellular processes. We have mainly concentrated on three different modular domains, namely SH2, SH3, and PDZ. The sequence and structural organization of the modular domains, their ligand binding specificity, and their molecular and cellular functions have been thoroughly described in this chapter.

The Chapter 2 is initiated with the description of various high-throughput techniques that are being widely used for determining the binding specificity of modular protein domains. Several computational methods have been developed that use these large-scale data for training their prediction models. We have described these computational methods in the second part of the chapter. Other computational methods that do not use the large-scale data have also been discussed. Although there are several methods available to predict modular domain-peptide interactions, they have several drawbacks, such as restrictive modeling assumption, limited coverage, pre-alignment problem etc., which can severely affect the prediction accuracy. Therefore, we have also highlighted these problems for better understanding the existing limitations for the computational prediction of the modular domain-peptide interactions.

In Chapter 3, we introduced efficient methods for predicting modular domain mediated interactions for three diverse modular domains, i.e., SH2, SH3, and PDZ. In addition, we showed how to tackle the various problems, which have been discussed in Chapter 2. We were able to handle the following major problems: (i) data-imbalance problem, (ii) model linearity, (iii) initial alignment of proline-rich peptides, (iv) generation of high-confidence negative data, and (v) lack of domain coverage.

In our study, we have used large-scale data that was derived from various high-throughput techniques, such as peptide array, microarray, and phage display [44, 58, 97, 164, 165, 191].

Although the large-scale data from these high-throughput techniques seemed to be a perfect data source for training the models, they have at least two types of problems: (i) a significant noise component and (ii) data imbalance between confirmed interactions (positive data) and experimentally proven non-interactions (negative data). For example, domain-peptide interaction data derived from microarray experiments contain both positive and negative data while the data derived from peptide array and phage display experiments contain only positive data. Furthermore, the interaction data may be rich for some domains while for other domains, either it is less or completely missing. To circumvent this limitation, we have taken two different strategies: (i) when at least a few negative examples are available, we have proposed an iterative data re-balancing strategy by introducing a semi-supervised learning approach. We have used this strategy for balancing the SH2 and PDZ domain-peptide interactions; and (ii) when there is no negative data available, in case of SH3 domain-peptide interactions, we have used a generative approach to balance the training dataset. Note that these techniques balance the training data by generating high-confidence negative data and thus are able to reduce the noise (false negatives) from the input data.

One of the major problems of existing approaches is that they are essentially linear models, which are not capable of handling complex interaction patterns. To overcome this limitation, we employed efficient kernel functions ranging from polynomial, to Gaussian, to advanced graph kernels, which can model complex interaction pattern by exploiting the positional dependency between the amino acids in the binding peptide.

An optimal alignment of proline-rich peptides, targeted by the SH3 domains, is a hard task. A minor error in the peptide alignment step can severely affect the performance of the predictive models (see Section 2.3.1). All the PWM-based methods rely on an initial peptide alignment for predicting SH3 domain mediated interactions and therefore produce suboptimal models. We have eliminated the need for an error-prone initial peptide alignment by introducing an advanced graph kernel approach.

In SH3-peptide interaction prediction, we have shown that the multi-domain models perform better than the single domain models. We extended this strategy for predicting PDZ domain mediated interactions across several organisms. In order to make a single model for multiple PDZ domains, we have clustered all the available PDZ domains from human, mouse, fly, and worm according to their binding specificity. Here, we have shown how we can leverage the information contained in related domains by building a single comprehensive model for a set of multiple modular domains. This strategy led us to provide the largest number of prediction models for PDZ domain to date. Moreover, these multi-domain models are easily applicable to the alignable novel domains where no training data is available.

In Chapter 4, all the applications and predictive performances of our three different methods have been reported. We have compared our methods with existing methods and achieved a much better performance in all three cases. The predictive performance was measured in terms of sensitivity, specificity, precision, AUC PR, and AUC ROC. Additionally, we have tested our all three methods on several manually curated databases of experimentally validated domain-peptide interactions and also achieved a better performance than the other existing approaches. At the end of this chapter, we have performed a genome-wide prediction for SH2, SH3, and PDZ domains to unveil the novel and biological insightful interactions. We have made all the top predictions freely available to the research community.

In Chapter 5, we have integrated our all three methods, i.e., SH2PepInt, SH3PepInt, and PDZPepInt, and developed a new and easy-to-use web server, namely MoDPepInt, for predicting modular domain-peptide interactions. The MoDPepInt web server has two different modes: (i) non-expert and (ii) expert mode. For non-expert use, a meta-web server has been implemented where users do not need to set any parameters. In this mode, the jobs are simultaneously submitted to all three tools and the top five predictions for each domain are reported. In the expert mode, the detailed parameter settings are available, however, user can easily access this mode from the non-expert mode. We believe biologists will be benefited by this web server and it will be very useful to pursue their research in this field.

In summary, this thesis has introduced a framework for the prediction of modular domain mediated interactions using high-throughput experimental data, which was applied to three diverse PRM families (i.e., SH2, SH3, and PDZ domain). Importantly, in this thesis, we have shown how to tackle the major computational problems to identify the modular domain mediated interactions. Finally, we introduce a new and efficient web server, namely MoDPepInt, which contains largest number of models to date for predicting modular domain-peptide interactions.

# Appendix A

# Supplementary material

## A.1. SH2 domain data

**Figure A.1.1.:** Averaged AUC ROC and AUC PR achieved by random train-test splitting method. (A, B) Showing the AUC ROC and AUC PR for the SVM performance, respectively. We achieved AUC ROC 0.9 and AUC PR 0.96. The figure is taken from the supplementary materials of [P4].

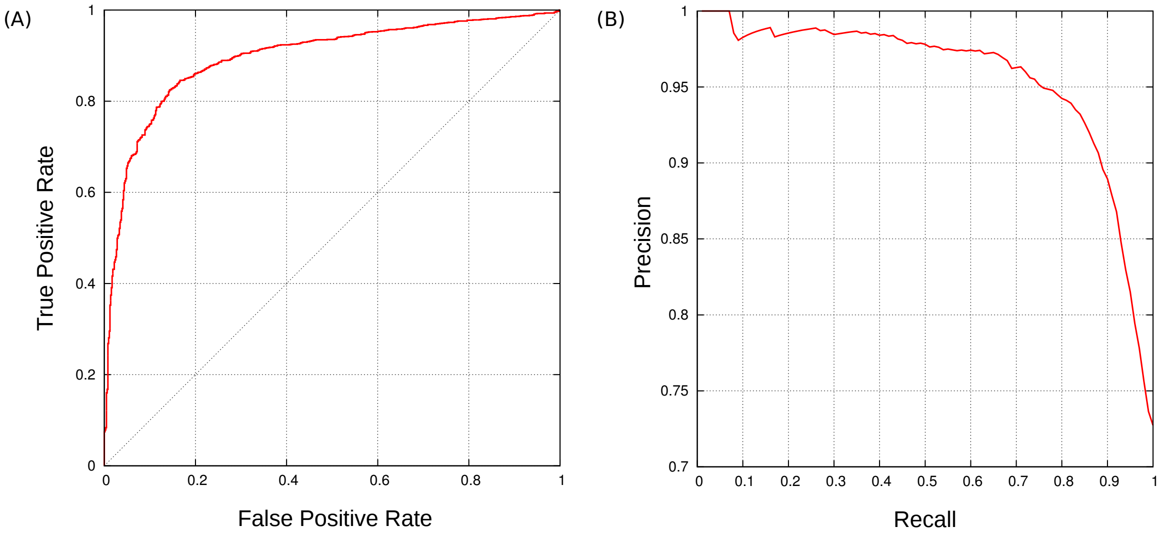


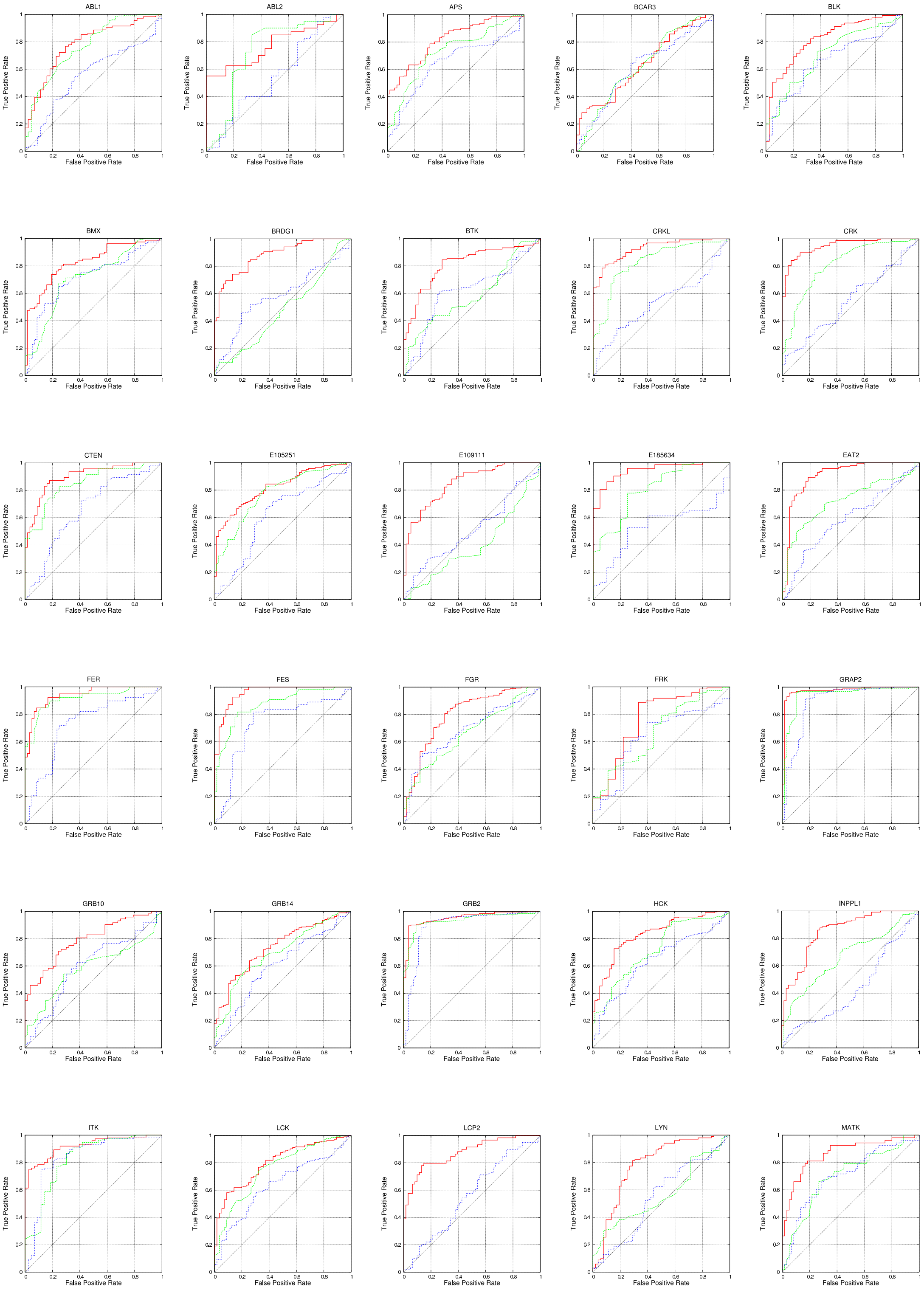
**Table A.1.1.:** Imbalanced level for confirmed presence or absence of peptide interactions for 51 SH2 domains. #int is the total number of interactions, #pos is the total number of positive interactions, #neg is the total number of negative interactions, and ratio of the positive and negative interactions. The table is taken from the supplementary materials of [P4].

| SH2 domain | # Interaction | # Positive | # Negative | Ratio |
|---|---|---|---|---|
| ABL1 | 222 | 178 | 44 | 4:1 |
| ABL2 | 61 | 40 | 21 | 2:1 |
| APS | 194 | 136 | 58 | 2:1 |
| BCAR3 | 145 | 92 | 53 | 2:1 |
| BLK | 278 | 238 | 40 | 6:1 |
| BMX | 137 | 80 | 57 | 1:1 |
| BRDG1 | 146 | 85 | 61 | 1:1 |
| BTK | 160 | 103 | 57 | 2:1 |
| CRKL | 177 | 131 | 46 | 3:1 |
| CRK | 204 | 158 | 46 | 3:1 |
| CTEN | 103 | 47 | 56 | 1:1 |
| E105251 | 204 | 143 | 61 | 2:1 |
| E109111 | 156 | 99 | 57 | 1:1 |
| E185634 | 93 | 73 | 20 | 4:1 |
| EAT2 | 200 | 141 | 59 | 2:1 |
| FER | 99 | 39 | 60 | 1:2 |
| FES | 115 | 55 | 60 | 1:1 |
| FGR | 328 | 278 | 50 | 6:1 |
| FRK | 284 | 266 | 18 | 15:1 |
| GRAP2 | 223 | 164 | 59 | 3:1 |
| GRB10 | 126 | 73 | 53 | 1:1 |
| GRB14 | 243 | 185 | 58 | 3:1 |
| GRB2 | 247 | 193 | 54 | 3:1 |
| HCK | 275 | 218 | 57 | 4:1 |
| INPPL1 | 184 | 123 | 61 | 2:1 |
| ITK | 120 | 77 | 43 | 2:1 |
| LCK | 273 | 217 | 56 | 4:1 |
| LCP2 | 120 | 59 | 61 | 1:1 |
| LYN | 154 | 102 | 52 | 2:1 |
| MATK | 113 | 53 | 60 | 1:1 |
| MIST | 93 | 83 | 10 | 8:1 |
| NCK1 | 160 | 109 | 51 | 2:1 |
| NCK2 | 149 | 101 | 48 | 2:1 |
| PTK6 | 266 | 206 | 60 | 3:1 |
| SH2B | 237 | 182 | 55 | 4:1 |
| SH2D1A | 394 | 337 | 57 | 6:1 |
| SH2D2A | 172 | 112 | 60 | 2:1 |
| SH2D3C | 130 | 76 | 54 | 1:1 |
| SHC1 | 202 | 151 | 51 | 3:1 |
| SHC3 | 114 | 58 | 56 | 1:1 |
| SOCS2 | 116 | 96 | 20 | 5:1 |
| SOCS5 | 80 | 70 | 10 | 7:1 |
| SRC | 373 | 333 | 40 | 8:1 |
| TEC | 214 | 165 | 49 | 3:1 |
| TENC1 | 252 | 197 | 55 | 4:1 |
| TENS1 | 177 | 124 | 53 | 2:1 |
| TNS | 261 | 205 | 56 | 4:1 |
| TXK | 188 | 133 | 55 | 2:1 |
| VAV1 | 115 | 59 | 56 | 1:1 |
| VAV2 | 89 | 40 | 49 | 1:1 |
| YES1 | 149 | 109 | 40 | 3:1 |

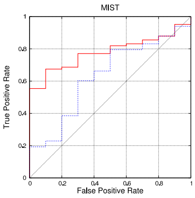
**Table A.1.2.:** Comparison of linear and non-linear kernel. We compare the AUC PR and AUC ROC of linear and non-linear kernel for each SH2 domain. The better performers are in bold. The table indicates that the non-linear (i.e., polynomial in our case) kernel performs better than the linear kernel. The table is taken from the supplementary materials of [P4].

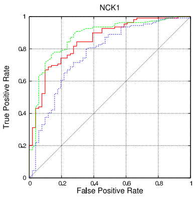
| Domain | AUC PR | | AUC ROC | |
|---|---|---|---|---|
| | Linear | Non-linear | Linear | Non-linear |
| ABL1 | 0.916 | **0.934** | 0.757 | **0.781** |
| ABL2 | 0.885 | **0.914** | 0.761 | **0.798** |
| APS | 0.909 | **0.923** | 0.801 | **0.823** |
| BCAR3 | 0.785 | **0.786** | **0.642** | 0.636 |
| BLK | 0.965 | **0.972** | 0.833 | **0.857** |
| BMX | 0.885 | **0.912** | 0.834 | **0.859** |
| BRDG1 | 0.925 | **0.942** | 0.886 | **0.907** |
| BTK | 0.872 | **0.906** | 0.773 | **0.826** |
| CRK | 0.982 | **0.985** | 0.943 | **0.947** |
| CRKL | 0.970 | **0.976** | 0.921 | **0.931** |
| CTEN | 0.841 | **0.910** | 0.865 | **0.903** |
| E105251 | 0.923 | **0.926** | 0.824 | **0.825** |
| E109111 | 0.903 | **0.912** | **0.855** | 0.846 |
| E185634 | **0.988** | 0.985 | **0.954** | 0.940 |
| EAT2 | 0.944 | **0.953** | 0.895 | **0.918** |
| FER | 0.874 | **0.928** | 0.914 | **0.956** |
| FES | 0.953 | **0.966** | 0.958 | **0.970** |
| FGR | 0.948 | **0.959** | 0.802 | **0.820** |
| FRK | 0.976 | 0.976 | 0.761 | **0.767** |
| GRAP2 | 0.981 | **0.987** | 0.961 | **0.972** |
| GRB10 | 0.845 | **0.879** | 0.783 | **0.808** |
| GRB14 | 0.878 | **0.905** | 0.710 | **0.739** |
| GRB2 | 0.979 | **0.987** | 0.937 | **0.956** |
| HCK | 0.939 | **0.952** | 0.810 | **0.838** |
| INPPL1 | 0.902 | **0.922** | 0.835 | **0.857** |
| ITK | 0.955 | **0.961** | 0.903 | **0.919** |
| LCK | **0.947** | 0.943 | **0.822** | 0.804 |
| LCP2 | 0.872 | **0.892** | 0.851 | **0.879** |
| LYN | 0.856 | **0.890** | 0.792 | **0.825** |
| MATK | **0.870** | 0.846 | **0.868** | 0.846 |
| MIST | **0.974** | 0.966 | **0.788** | 0.739 |
| NCK1 | 0.903 | **0.923** | 0.818 | **0.853** |
| NCK2 | 0.924 | **0.949** | 0.857 | **0.894** |
| PTK6 | 0.909 | **0.935** | 0.771 | **0.803** |
| SH2B | 0.934 | **0.939** | 0.804 | **0.824** |
| SH2D1A | 0.931 | **0.938** | 0.725 | **0.737** |
| SH2D2A | 0.843 | **0.878** | 0.742 | **0.777** |
| SH2D3C | 0.865 | **0.887** | 0.825 | **0.832** |
| SHC1 | 0.902 | **0.915** | 0.763 | **0.781** |
| SHC3 | 0.866 | **0.870** | **0.866** | 0.855 |
| SOCS2 | 0.969 | **0.980** | 0.873 | **0.915** |
| SOCS5 | 0.989 | **0.991** | 0.921 | **0.936** |
| SRC | 0.959 | 0.959 | **0.756** | 0.740 |
| TEC | 0.907 | **0.922** | 0.781 | **0.791** |
| TENC1 | 0.930 | **0.941** | 0.794 | **0.810** |
| TENS1 | 0.923 | **0.935** | 0.842 | **0.852** |
| TNS | 0.938 | **0.954** | 0.807 | **0.848** |
| TXK | 0.889 | **0.905** | 0.784 | **0.793** |
| VAV1 | 0.946 | **0.947** | **0.942** | 0.930 |
| VAV2 | 0.902 | **0.903** | 0.891 | **0.904** |
| YES1 | 0.885 | **0.924** | 0.721 | **0.800** |

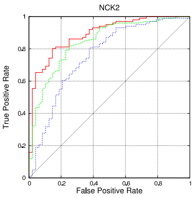
**Figure A.1.2.:** AUC ROC comparison. AUC ROC curves achieved by SVM (red lines), `SMALI` (green dashed lines), and energy model (blue dotted lines) for each SH2 domain. The figure is taken from the supplementary materials of [P4].

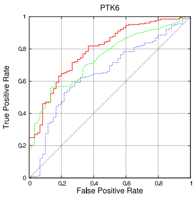


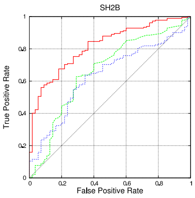
## A.1. SH2 domain data

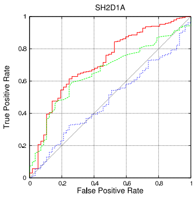


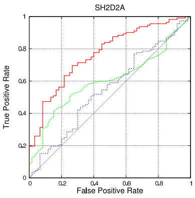


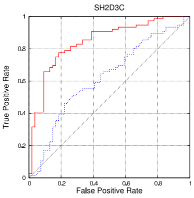


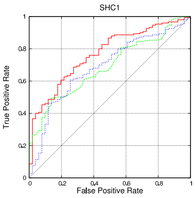


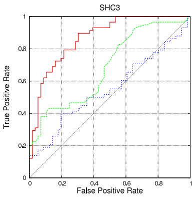


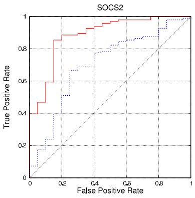


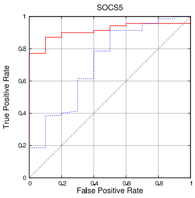


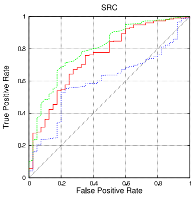


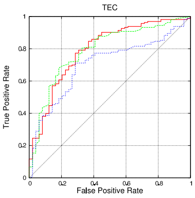


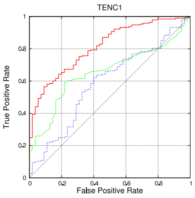


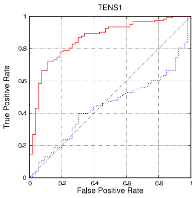


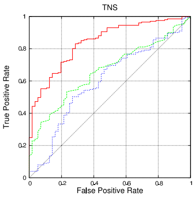


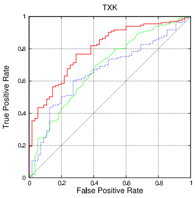


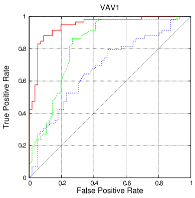


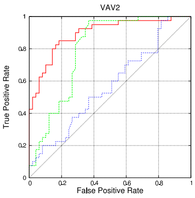












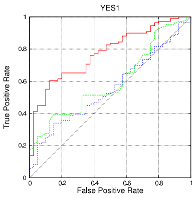
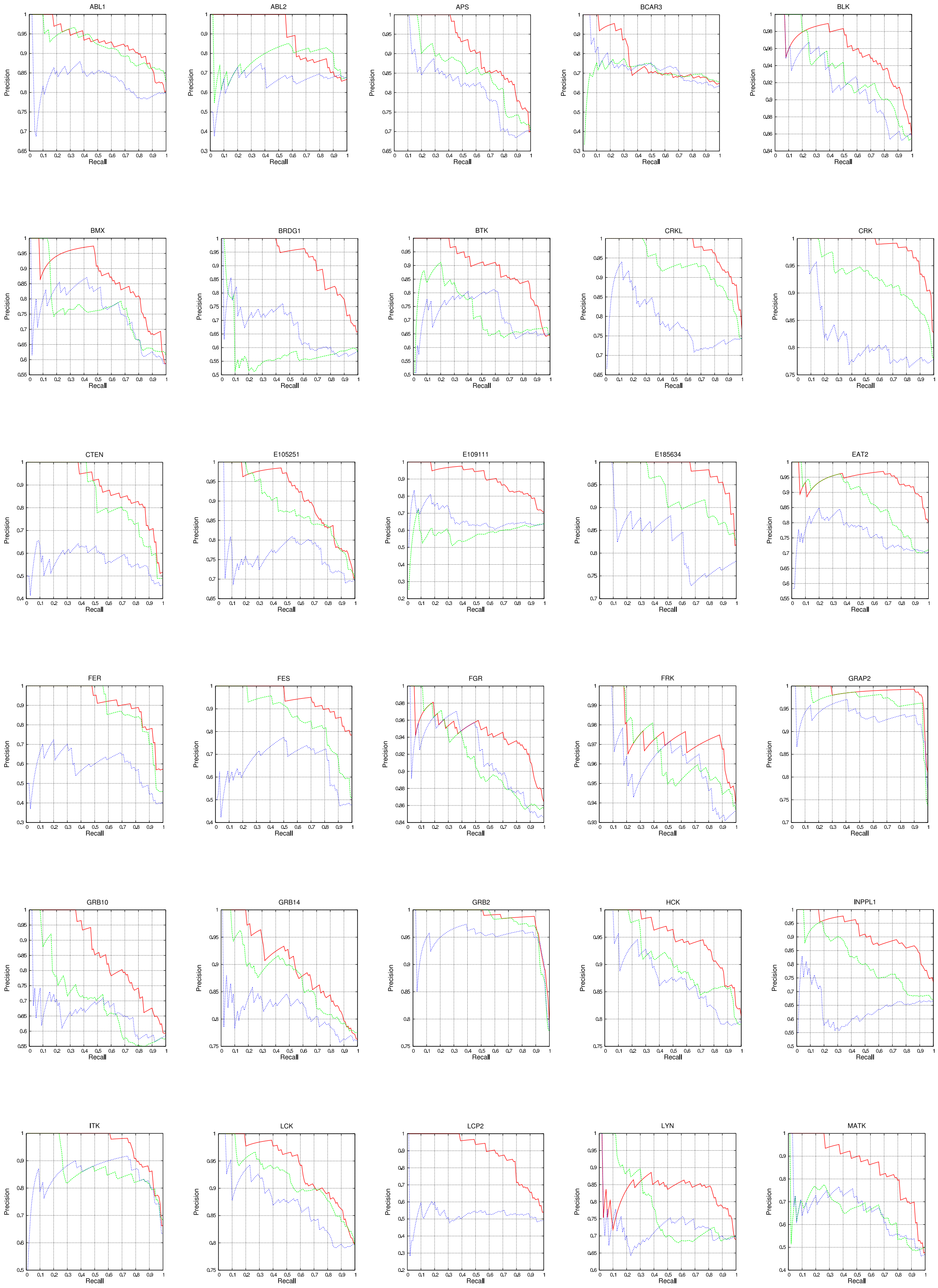
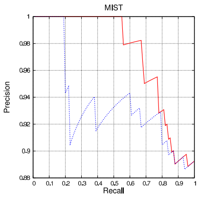
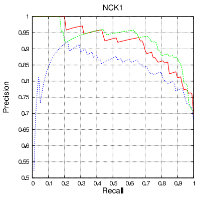
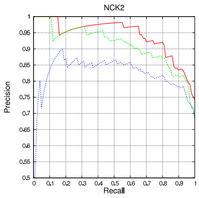


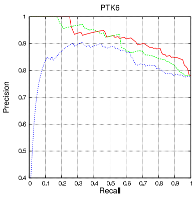
**Figure A.1.3.:** AUC PR comparison. AUC PR curves achieved by SVM (red lines), `SMALI` (green dashed lines ), and energy model (blue dotted lines) for each SH2 domain. The figure is taken from the supplementary materials of [P4].

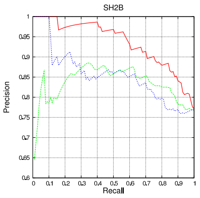


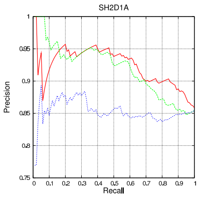

MIST
1
0.98
0.96
0.94
0.92
0.9
0.88
Precision
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1
Recall

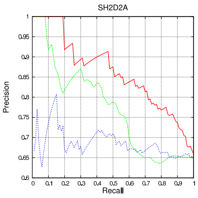

NCK1
1
0.95
0.9
0.85
0.8
0.75
0.7
0.65
0.6
0.55
0.5
Precision
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1
Recall

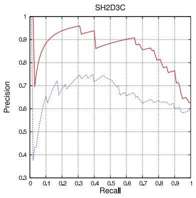

NCK2
1
0.95
0.9
0.85
0.8
0.75
0.7
0.65
0.6
0.55
0.5
Precision
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1
Recall

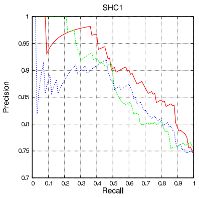

PTK6
1
0.9
0.8
0.7
0.6
0.5
0.4
Precision
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1
Recall

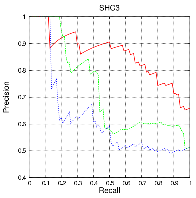

SH2B
1
0.95
0.9
0.85
0.8
0.75
0.7
0.65
0.6
Precision
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1
Recall

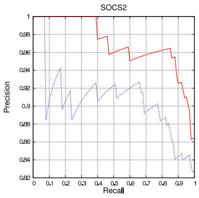

SH2D1A
1
0.95
0.9
0.85
0.8
0.75
Precision
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1
Recall

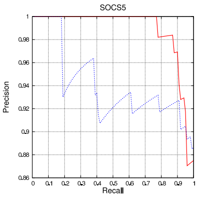

SH2D2A
1
0.95
0.9
0.85
0.8
0.75
0.7
0.65
0.6
Precision
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1
Recall

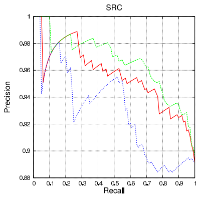

SH2D3C
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
Precision
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1
Recall

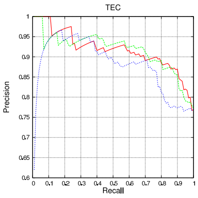

SHC1
1
0.95
0.9
0.85
0.8
0.75
0.7
Precision
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1
Recall

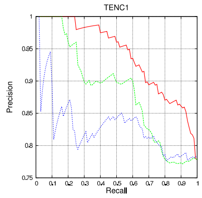

SHC3
1
0.9
0.8
0.7
0.6
0.5
0.4
Precision
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1
Recall

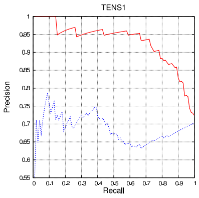

SOCS2
1
0.98
0.96
0.94
0.92
0.9
0.88
0.86
0.84
0.82
Precision
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1
Recall

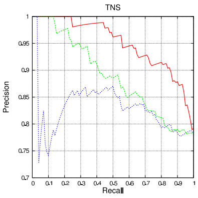

SOCS5
1
0.98
0.96
0.94
0.92
0.9
0.88
0.86
Precision
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1
Recall

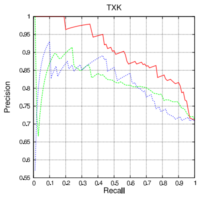

SRC
1
0.98
0.96
0.94
0.92
0.9
0.88
Precision
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1
Recall

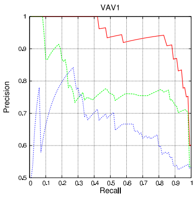

TEC
1
0.95
0.9
0.85
0.8
0.75
0.7
0.65
0.6
Precision
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1
Recall

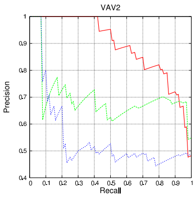

TENC1
1
0.95
0.9
0.85
0.8
0.75
Precision
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1
Recall


TENS1
1
0.95
0.9
0.85
0.8
0.75
0.7
0.65
0.6
0.55
Precision
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1
Recall


TNS
1
0.95
0.9
0.85
0.8
0.75
0.7
Precision
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1
Recall


TXK
1
0.95
0.9
0.85
0.8
0.75
0.7
0.65
0.6
0.55
Precision
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1
Recall


VAV1
1
0.9
0.8
0.7
0.6
0.5
Precision
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1
Recall


VAV2
1
0.9
0.8
0.7
0.6
0.5
0.4
Precision
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1
Recall

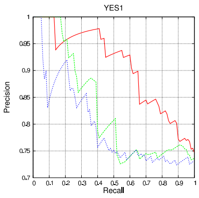
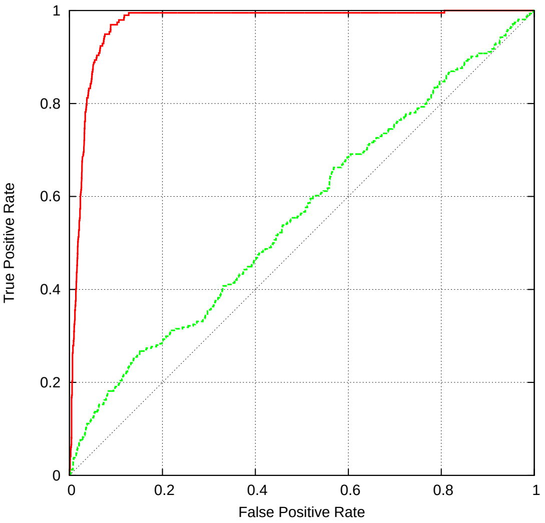

YES1
1
0.95
0.9
0.85
0.8
0.75
0.7
Precision
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1
Recall

**Figure A.1.4.:** Binding and non-binding energy comparison with different microarray data. AUC ROC of the dataset II and dataset III derived by the energy model [230]. Indicating the AUR ROC of the experiments and clearly showing the AUR ROC of dataset II, 0.97 (red line) is much higher than AUR ROC of dataset III, 0.56 (green dashed line). This result is probably due to some over-training issues. The figure is taken from the supplementary materials of [P4].



## A.2. SH3 domain data

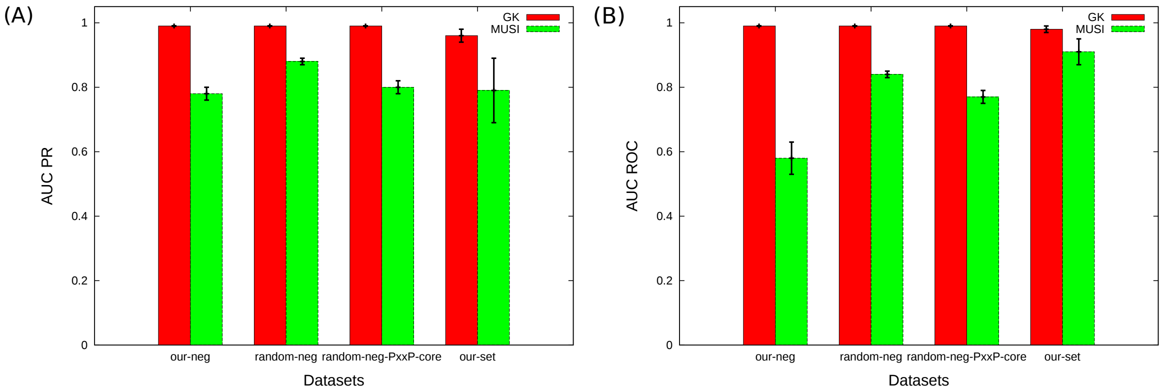


**Figure A.2.1.:** Precision-recall curves and AUC ROC curves for the Single Domain Graph kernel (GK) and the `MUSI` tool for different datasets of SRC SH3 domain. The error bars represent respective standard deviation. This figure clearly shows the GK performs much better than the `MUSI` tool. The figure is taken from the supplementary materials of [P3].

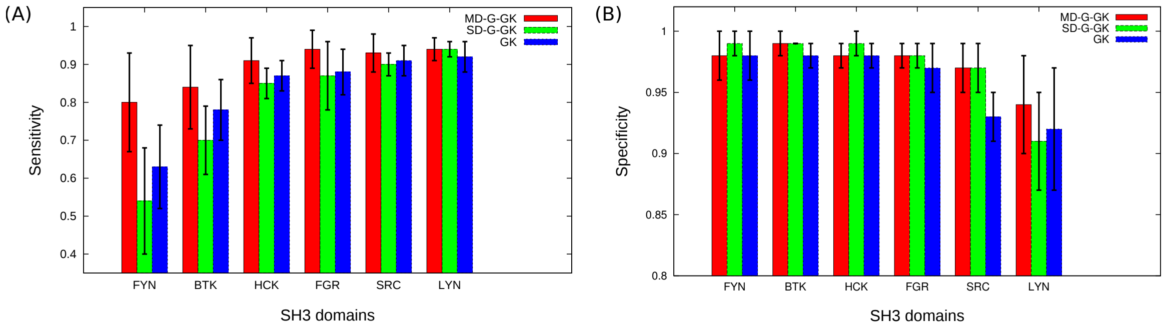


**Figure A.2.2.:** Sensitivity and specificity for the Multi-Domain Gaussian Graph Kernel (MD-G-GK), the Single Domain Gaussian Graph kernel (SD-G-GK), and the Single Domain Linear Graph Kernel (GK) for 6 related SH3 domains. The error bars represent respective standard deviation. The figure is adapted from the supplementary materials of [P3].

**Table A.2.1.:** Experimental data derived from [58]. Here, we report the binding specificity of each domains. We found seven peptide motifs to describe the whole dataset. Majority of the human SH3 domains found to be bound with class I and/or class II peptides. The numbers represent the sharing percentage value of a motif with respective SH3 domains. Numbers in parenthesis indicate the sharing percentage value of a peptide motif after removing all class I and class II peptides. The table is taken from the supplementary materials of [P3].

| **SH3 domain** | **# Positive** | **Class I (%)** | **Class II (%)** | **Class I & II (%)** | **PxRP (%)** | **PxxPR (%)** | **PxxDY (%)** | **RxxKP (%)** | **PPPPP (%)** |
|---|---|---|---|---|---|---|---|---|---|
| ABL-61-121 | 202 | 40.59 | 36.63 | 7.92 | 15.84(6.44) | 8.91(1.98) | 0(0) | 3.47(1.49) | 15.84(6.44) |
| AMPHIPHYSINI-622-695 | 322 | 20.19 | 50 | 9.63 | 49.38(36.02) | 34.47(10.87) | 0(0) | 3.42(1.55) | 2.48(0.93) |
| AMPHIPHYSINII-520-592 | 215 | 20 | 65.12 | 12.09 | 22.79(16.74) | 49.3(14.42) | 0(0) | 4.19(0.93) | 4.19(1.4) |
| ARGBP1-451-510 | 134 | 31.34 | 52.24 | 11.94 | 16.42(9.7) | 10.45(5.97) | 0(0) | 6.72(5.22) | 2.99(2.24) |
| ARGBP2a-1041-1100 | 72 | 29.17 | 30.56 | 8.33 | 52.78(30.56) | 13.89(4.17) | 0(0) | 4.17(2.78) | 12.5(6.94) |
| ARGBP2a-863-922 | 321 | 29.28 | 40.81 | 6.23 | 40.5(19.31) | 18.07(3.43) | 0(0) | 4.05(2.8) | 9.66(4.98) |
| ARGBP2a-938-999 | 307 | 37.79 | 59.28 | 13.68 | 11.4(6.51) | 23.78(2.61) | 0(0) | 4.56(2.61) | 5.86(0.33) |
| ARHGAP12-12-74 | 122 | 40.98 | 32.79 | 6.56 | 18.85(8.2) | 13.11(2.46) | 0(0) | 2.46(1.64) | 21.31(6.56) |
| ARHGEF16-629-689 | 88 | 34.09 | 39.77 | 10.23 | 15.91(12.5) | 10.23(4.55) | 0(0) | 13.64(9.09) | 0(0) |
| BOG25-55-114 | 368 | 34.24 | 33.7 | 5.43 | 11.41(8.42) | 7.07(2.45) | 0.82(0.82) | 3.53(2.45) | 2.99(1.36) |
| BTK-214-274 | 264 | 61.36 | 37.88 | 10.61 | 16.67(6.06) | 18.18(1.52) | 0(0) | 1.52(0.38) | 6.82(1.52) |
| CAP-1049-1108 | 133 | 24.06 | 39.1 | 6.02 | 45.86(21.8) | 23.31(2.26) | 0.75(0.75) | 3.76(0.75) | 16.54(8.27) |
| CAP-1123-1184 | 92 | 29.35 | 43.48 | 4.35 | 23.91(14.13) | 9.78(1.09) | 0(0) | 4.35(2.17) | 7.61(4.35) |
| CIN85-1-58 | 130 | 21.54 | 39.23 | 12.31 | 20(13.08) | 55.38(37.69) | 0(0) | 4.62(1.54) | 0.77(0.77) |
| CIN85-267-328 | 305 | 21.64 | 47.54 | 12.79 | 16.39(10.49) | 62.62(38.36) | 0(0) | 3.93(2.95) | 0.98(0.33) |
| COOL1-P85-184-243 | 353 | 32.86 | 45.61 | 9.07 | 20.4(10.76) | 18.41(4.82) | 0.85(0.85) | 4.53(2.83) | 3.12(0.85) |
| CSK-9-70 | 43 | 41.86 | 44.19 | 6.98 | 9.3(2.33) | 6.98(0) | 0(0) | 2.33(0) | 4.65(0) |
| DDEF2-944-1006 | 336 | 32.74 | 34.52 | 11.01 | 32.14(25.89) | 21.13(10.42) | 0.3(0.3) | 8.63(5.95) | 4.46(2.08) |
| DLG1-581-651 | 42 | 47.62 | 21.43 | 4.76 | 4.76(4.76) | 4.76(0) | 0(0) | 7.14(7.14) | 2.38(2.38) |
| DOCK1-9-70 | 323 | 35.6 | 32.2 | 7.12 | 19.81(13.62) | 11.15(3.41) | 0.31(0.31) | 5.88(4.33) | 6.19(3.41) |
| ENDOPHB1-305-365 | 122 | 39.34 | 37.7 | 8.2 | 10.66(9.84) | 9.02(4.1) | 0(0) | 6.56(4.92) | 0(0) |
| ENDOPHILIN1-290-349 | 181 | 31.49 | 38.67 | 8.84 | 55.8(31.49) | 19.34(7.18) | 0(0) | 3.31(2.21) | 6.63(1.66) |
| ENDOPHILIN2-306-365 | 387 | 37.98 | 37.47 | 11.89 | 37.47(23.26) | 16.28(5.68) | 0(0) | 5.43(4.13) | 3.88(0.78) |
| ENDOPHILIN3-285-344 | 357 | 35.29 | 36.41 | 10.08 | 47.62(27.73) | 17.93(6.44) | 0(0) | 4.76(3.64) | 3.64(1.12) |
| EPS8-531-590 | 48 | 18.75 | 6.25 | 2.08 | 2.08(2.08) | 2.08(2.08) | 68.75(64.58) | 0(0) | 8.33(8.33) |
| EPS8L2-492-551 | 236 | 31.36 | 22.03 | 4.66 | 13.14(9.75) | 5.08(2.12) | 5.93(5.08) | 3.39(2.54) | 2.12(2.12) |

*Continued on next page .....*

**Table A.2.1** – *Continued from previous page*

| SH3 domain | # Positive | Class I<br>(%) | Class II<br>(%) | Class I & II<br>(%) | PxRP<br>(%) | PxxPR<br>(%) | PxxDY<br>(%) | RxxKP<br>(%) | PPPPP<br>(%) |
|---|---|---|---|---|---|---|---|---|---|
| FGR-77-138 | 286 | 62.59 | 40.91 | 12.59 | 17.48(5.94) | 17.83(1.05) | 0(0) | 3.15(0.35) | 6.64(1.05) |
| FISH-166-225 | 31 | 41.94 | 41.94 | 16.13 | 35.48(12.9) | 19.35(0) | 0(0) | 3.23(0) | 22.58(9.68) |
| FNBP1L-538-599 | 188 | 42.02 | 40.96 | 10.64 | 21.81(7.45) | 14.36(2.66) | 0(0) | 2.66(1.6) | 28.19(11.7) |
| FRK-42-110 | 281 | 40.21 | 43.77 | 8.9 | 16.37(7.47) | 15.66(1.78) | 0(0) | 2.85(1.42) | 6.41(2.49) |
| FYN-82-143 | 187 | 62.57 | 37.43 | 10.7 | 15.51(3.74) | 16.04(1.07) | 0(0) | 3.21(0) | 4.81(1.6) |
| GRAP2-1-56 | 72 | 37.5 | 37.5 | 9.72 | 12.5(11.11) | 6.94(2.78) | 0(0) | 13.89(9.72) | 0(0) |
| GRAP2-271-330 | 219 | 28.31 | 40.18 | 11.42 | 28.31(19.18) | 18.72(5.48) | 0(0) | 19.18(15.98) | 4.11(2.28) |
| GRB2-156-215 | 217 | 44.24 | 56.68 | 10.6 | 13.36(3.23) | 38.71(2.76) | 0.46(0) | 3.69(0.46) | 10.14(2.76) |
| GRB2-1-58 | 213 | 32.39 | 38.97 | 12.21 | 68.54(37.09) | 26.76(8.92) | 0(0) | 4.23(3.29) | 7.51(1.41) |
| HCK-78-138 | 392 | 46.94 | 53.57 | 10.97 | 16.58(4.08) | 18.11(1.28) | 0(0) | 1.79(0.51) | 4.08(1.53) |
| INTERSECTIN1-1002-1060 | 263 | 44.49 | 61.98 | 14.45 | 34.6(7.22) | 34.6(1.52) | 0(0) | 2.66(0.38) | 4.94(0) |
| INTERSECTIN1-1074-1138 | 188 | 31.91 | 29.26 | 5.32 | 11.17(10.11) | 7.45(3.72) | 1.06(1.06) | 3.72(3.19) | 0(0) |
| INTERSECTIN1-1155-1214 | 108 | 24.07 | 71.3 | 15.74 | 24.07(10.19) | 40.74(3.7) | 0(0) | 3.7(0) | 10.19(3.7) |
| INTERSECTIN1-740-806 | 239 | 32.22 | 69.87 | 13.39 | 21.76(5.44) | 31.38(1.26) | 0(0) | 2.93(0.84) | 6.69(1.67) |
| INTERSECTIN1-913-971 | 74 | 41.89 | 31.08 | 4.05 | 10.81(5.41) | 6.76(2.7) | 0(0) | 2.7(1.35) | 10.81(6.76) |
| IRSP53-374-437 | 362 | 32.87 | 36.46 | 5.8 | 14.64(7.46) | 9.94(2.49) | 0.55(0.55) | 4.14(3.04) | 9.67(6.63) |
| LCK-61-121 | 150 | 49.33 | 52.67 | 12 | 22(4.67) | 25.33(2.67) | 0(0) | 2.67(0) | 5.33(1.33) |
| LYN-63-123 | 737 | 41.66 | 50.34 | 8.96 | 18.18(5.56) | 19.4(1.63) | 0.14(0.14) | 2.04(1.09) | 6.51(2.04) |
| MLK3-41-105 | 131 | 28.24 | 60.31 | 12.21 | 14.5(8.4) | 29.01(4.58) | 0(0) | 9.16(4.58) | 6.11(2.29) |
| MPP1-158-228 | 142 | 38.73 | 33.8 | 7.75 | 16.9(13.38) | 11.27(4.23) | 0(0) | 8.45(5.63) | 2.11(0.7) |
| MYO7A-1603-1672 | 156 | 36.54 | 30.13 | 4.49 | 13.46(10.9) | 8.33(3.85) | 0.64(0.64) | 5.77(4.49) | 0.64(0.64) |
| NCF1-226-285 | 94 | 36.17 | 41.49 | 9.57 | 15.96(11.7) | 10.64(3.19) | 0(0) | 12.77(7.45) | 2.13(0) |
| NCK2-195-257 | 137 | 35.04 | 48.18 | 7.3 | 18.98(13.14) | 16.79(3.65) | 0(0) | 3.65(2.19) | 1.46(0) |
| NPHP1-152-212 | 308 | 72.4 | 22.08 | 11.04 | 28.9(15.58) | 11.36(1.95) | 0(0) | 0.32(0) | 6.82(0.65) |
| N-SRC-84-145 | 181 | 44.75 | 38.67 | 9.94 | 14.92(7.73) | 14.92(2.76) | 0(0) | 3.31(2.21) | 2.76(1.66) |
| OSTF1-12-71 | 437 | 39.59 | 72.77 | 18.54 | 18.76(2.52) | 37.07(1.37) | 0(0) | 2.75(0.92) | 16.02(2.06) |
| P51NOX-399-458 | 73 | 32.88 | 32.88 | 6.85 | 13.7(13.7) | 4.11(2.74) | 0(0) | 15.07(12.33) | 1.37(1.37) |
| PAC2-426-486 | 447 | 48.55 | 46.53 | 17.9 | 25.06(14.09) | 20.13(4.03) | 0.22(0.22) | 4.47(2.68) | 5.82(0.45) |
| PAC3-363-424 | 260 | 52.69 | 40 | 14.62 | 26.15(13.85) | 15(2.69) | 0.77(0.77) | 2.31(1.15) | 7.31(1.15) |
| PIK3R1-3-79 | 393 | 48.35 | 55.22 | 13.49 | 34.61(11.45) | 24.43(1.78) | 0.25(0) | 1.78(0.51) | 3.82(0) |

*Continued on next page .....*

**Table A.2.1** – *Continued from previous page*

| SH3 domain | # Positive | Class I (%) | Class II (%) | Class I & II (%) | PxRP (%) | PxxPR (%) | PxxDY (%) | RxxKP (%) | PPPPP (%) |
|---|---|---|---|---|---|---|---|---|---|
| PLCG1-791-851 | 442 | 30.09 | 56.33 | 8.82 | 27.6(9.28) | 39.59(8.82) | 0(0) | 2.49(1.13) | 7.24(0.9) |
| RASGAP-279-341 | 315 | 35.87 | 35.24 | 8.57 | 19.37(15.24) | 13.02(4.44) | 0(0) | 8.57(6.98) | 1.59(0.63) |
| RIMB1-1625-1693 | 219 | 66.67 | 29.68 | 13.7 | 15.53(11.42) | 8.22(3.2) | 0(0) | 2.74(1.83) | 2.74(1.83) |
| RIMB1-1764-1831 | 147 | 74.15 | 25.85 | 12.93 | 17.69(11.56) | 7.48(1.36) | 0(0) | 2.04(1.36) | 4.08(2.72) |
| RUSC1-844-902 | 50 | 46 | 40 | 20 | 12(8) | 8(4) | 0(0) | 6(4) | 0(0) |
| SH3PX3-1-61 | 286 | 69.23 | 28.32 | 12.24 | 35.66(14.69) | 14.69(1.4) | 0.35(0) | 0(0) | 16.78(1.4) |
| SNX18-1-61 | 200 | 66 | 36.5 | 13.5 | 35.5(7) | 17(1) | 0(0) | 1.5(0.5) | 17(1.5) |
| SNX9-1-62 | 389 | 58.1 | 32.65 | 13.88 | 43.44(21.34) | 15.68(3.34) | 0(0) | 1.29(0.26) | 13.11(1.54) |
| SRC-84-145 | 527 | 50.47 | 46.68 | 11.2 | 17.65(4.17) | 21.25(1.33) | 0(0) | 2.47(0.57) | 9.3(2.66) |
| STAM1-210-269 | 259 | 39.38 | 47.1 | 17.37 | 22.39(13.51) | 16.99(4.25) | 0(0) | 12.36(10.04) | 2.32(0.77) |
| STAM2-202-261 | 28 | 32.14 | 64.29 | 21.43 | 7.14(0) | 14.29(0) | 0(0) | 21.43(10.71) | 3.57(3.57) |
| TUBA-145-204 | 187 | 51.34 | 44.92 | 17.11 | 24.6(13.37) | 21.39(3.74) | 0.53(0.53) | 2.67(0) | 8.02(0.53) |
| TUBA-1513-1576 | 414 | 24.88 | 42.75 | 5.31 | 29.47(17.87) | 14.73(3.62) | 0.24(0.24) | 0.97(0.48) | 9.66(5.56) |
| TUBA-2-61 | 100 | 30 | 42 | 8 | 39(22) | 10(4) | 0(0) | 4(2) | 4(2) |

**Table A.2.2.:** The total number of positive and negative data along with the ratio for each SH3 domain used in our study. The table is taken from the supplementary materials of [P3].

| SH3 domain | # Positive | # Negative | Ratio |
|---|---|---|---|
| ABL-61-121 | 202 | 1633 | 1:8 |
| AMPHIPHYSINI-622-695 | 322 | 2354 | 1:7 |
| AMPHIPHYSINII-520-592 | 215 | 2668 | 1:12 |
| ARGBP1-451-510 | 134 | 5295 | 1:39 |
| ARGBP2a-1041-1100 | 72 | 7680 | 1:106 |
| ARGBP2a-863-922 | 321 | 2174 | 1:6 |
| ARGBP2a-938-999 | 307 | 2949 | 1:9 |
| ARHGAP12-12-74 | 122 | 3199 | 1:26 |
| ARHGEF16-629-689 | 88 | 5114 | 1:58 |
| BOG25-55-114 | 368 | 3401 | 1:9 |
| BTK-214-274 | 264 | 2465 | 1:9 |
| CAP-1049-1108 | 133 | 2509 | 1:18 |
| CAP-1123-1184 | 92 | 6132 | 1:66 |
| CIN85-1-58 | 130 | 6739 | 1:51 |
| CIN85-267-328 | 305 | 3962 | 1:12 |
| COOL1-P85-184-243 | 353 | 1813 | 1:5 |
| CSK-9-70 | 43 | 7403 | 1:172 |
| DDEF2-944-1006 | 336 | 1482 | 1:4 |
| DLG1-581-651 | 42 | 8591 | 1:204 |
| DOCK1-9-70 | 323 | 1224 | 1:3 |
| ENDOPHB1-305-365 | 122 | 7430 | 1:60 |
| ENDOPHILIN1-290-349 | 181 | 1462 | 1:8 |
| ENDOPHILIN2-306-365 | 387 | 3657 | 1:9 |
| ENDOPHILIN3-285-344 | 357 | 2589 | 1:7 |
| EPS8-531-590 | 48 | 9048 | 1:188 |
| EPS8L2-492-551 | 236 | 5991 | 1:25 |
| FGR-77-138 | 286 | 799 | 1:2 |
| FISH-166-225 | 31 | 6918 | 1:223 |
| FNBP1L-538-599 | 188 | 2256 | 1:12 |
| FRK-42-110 | 281 | 2571 | 1:9 |
| FYN-82-143 | 187 | 1650 | 1:8 |
| GRAP2-1-56 | 72 | 7640 | 1:106 |
| GRAP2-271-330 | 219 | 3586 | 1:16 |
| GRB2-156-215 | 217 | 2017 | 1:9 |
| GRB2-1-58 | 213 | 1171 | 1:5 |
| HCK-78-138 | 392 | 2105 | 1:5 |
| INTERSECTIN1-1002-1060 | 263 | 1522 | 1:5 |
| INTERSECTIN1-1074-1138 | 188 | 4301 | 1:22 |
| INTERSECTIN1-1155-1214 | 108 | 3401 | 1:31 |
| INTERSECTIN1-740-806 | 239 | 1679 | 1:7 |
| INTERSECTIN1-913-971 | 74 | 7453 | 1:100 |
| IRSP53-374-437 | 362 | 2540 | 1:7 |
| LCK-61-121 | 150 | 3993 | 1:26 |
| LYN-63-123 | 737 | 804 | 1:1 |
| MLK3-41-105 | 131 | 4683 | 1:35 |
| MPP1-158-228 | 142 | 4855 | 1:34 |

*Continued on next page .....*

**Table A.2.2** – *Continued from previous page*

| SH3 domain | # Positive | # Negative | Ratio |
|---|---|---|---|
| MYO7A-1603-1672 | 156 | 4795 | 1:30 |
| NCF1-226-285 | 94 | 6157 | 1:65 |
| NCK2-195-257 | 137 | 5112 | 1:37 |
| NPHP1-152-212 | 308 | 2770 | 1:8 |
| N-SRC-84-145 | 181 | 4219 | 1:23 |
| OSTF1-12-71 | 437 | 1747 | 1:3 |
| P51NOX-399-458 | 73 | 7732 | 1:105 |
| PAC2-426-486 | 447 | 1731 | 1:3 |
| PAC3-363-424 | 260 | 1354 | 1:5 |
| PIK3R1-3-79 | 393 | 1966 | 1:5 |
| PLCG1-791-851 | 442 | 1804 | 1:4 |
| RASGAP-279-341 | 315 | 3760 | 1:11 |
| RIMB1-1625-1693 | 219 | 4063 | 1:18 |
| RIMB1-1764-1831 | 147 | 6070 | 1:41 |
| RUSC1-844-902 | 50 | 8154 | 1:163 |
| SH3PX3-1-61 | 286 | 1219 | 1:4 |
| SNX18-1-61 | 200 | 1143 | 1:5 |
| SNX9-1-62 | 389 | 1546 | 1:3 |
| SRC-84-145 | 527 | 1268 | 1:2 |
| STAM1-210-269 | 259 | 3241 | 1:12 |
| STAM2-202-261 | 28 | 8775 | 1:313 |
| TUBA-145-204 | 187 | 1586 | 1:8 |
| TUBA-1513-1576 | 414 | 2346 | 1:5 |
| TUBA-2-61 | 100 | 5417 | 1:54 |

**Table A.2.3.:** The results achieved by the single domain graph kernel (GK) approach for each human SH3 domain. We report the average sensitivity, specificity, precision, AUC PR, and AUC ROC for each domain. The table is taken from the supplementary materials of [P3].

| SH3 domain | Sensitivity | Specificity | Precision | AUC PR | AUC ROC |
|---|---|---|---|---|---|
| ABL-61-121 | 0.63 | 0.98 | 0.78 | 0.76 | 0.93 |
| AMPHIPHYSINI-622-695 | 0.75 | 0.99 | 0.91 | 0.9 | 0.97 |
| AMPHIPHYSINII-520-592 | 0.64 | 0.99 | 0.85 | 0.81 | 0.94 |
| ARGBP1-451-510 | 0.49 | 1 | 0.77 | 0.69 | 0.95 |
| ARGBP2a-1041-1100 | 0.15 | 1 | 0.75 | 0.32 | 0.73 |
| ARGBP2a-863-922 | 0.73 | 0.97 | 0.78 | 0.83 | 0.95 |
| ARGBP2a-938-999 | 0.81 | 0.98 | 0.85 | 0.91 | 0.98 |
| ARHGAP12-12-74 | 0.26 | 0.99 | 0.6 | 0.54 | 0.92 |
| ARHGEF16-629-689 | 0.53 | 1 | 0.75 | 0.74 | 0.98 |
| BOG25-55-114 | 0.75 | 0.96 | 0.73 | 0.81 | 0.97 |
| BTK-214-274 | 0.78 | 0.98 | 0.8 | 0.88 | 0.97 |
| CAP-1049-1108 | 0.45 | 0.99 | 0.78 | 0.67 | 0.94 |
| CAP-1123-1184 | 0.08 | 1 | 0.43 | 0.26 | 0.82 |
| CIN85-1-58 | 0.38 | 1 | 0.87 | 0.6 | 0.94 |
| CIN85-267-328 | 0.87 | 0.99 | 0.84 | 0.93 | 0.99 |
| COOL1-P85-184-243 | 0.7 | 0.98 | 0.87 | 0.89 | 0.97 |
| CSK-9-70 | 0.04 | 1 | 0.2 | 0.26 | 0.87 |
| DDEF2-944-1006 | 0.81 | 0.97 | 0.88 | 0.91 | 0.96 |
| DLG1-581-651 | 0.04 | 1 | 0.15 | 0.26 | 0.93 |
| DOCK1-9-70 | 0.79 | 0.94 | 0.79 | 0.88 | 0.96 |
| ENDOPHB1-305-365 | 0.44 | 0.99 | 0.55 | 0.54 | 0.98 |
| ENDOPHILIN1-290-349 | 0.75 | 0.98 | 0.82 | 0.87 | 0.96 |
| ENDOPHILIN2-306-365 | 0.78 | 0.97 | 0.77 | 0.84 | 0.97 |
| ENDOPHILIN3-285-344 | 0.73 | 0.98 | 0.88 | 0.87 | 0.97 |
| EPS8-531-590 | 0.43 | 1 | 0.59 | 0.56 | 0.95 |
| EPS8L2-492-551 | 0.6 | 0.98 | 0.54 | 0.6 | 0.96 |
| FGR-77-138 | 0.88 | 0.97 | 0.91 | 0.96 | 0.98 |
| FISH-166-225 | 0 | 1 | 0 | 0.01 | 0.59 |
| FNBP1L-538-599 | 0.65 | 0.98 | 0.72 | 0.79 | 0.96 |
| FRK-42-110 | 0.71 | 0.99 | 0.87 | 0.87 | 0.96 |
| FYN-82-143 | 0.63 | 0.98 | 0.76 | 0.73 | 0.88 |
| GRAP2-1-56 | 0.19 | 1 | 0.46 | 0.4 | 0.96 |
| GRAP2-271-330 | 0.7 | 0.99 | 0.85 | 0.82 | 0.96 |
| GRB2-156-215 | 0.66 | 0.98 | 0.8 | 0.8 | 0.92 |
| GRB2-1-58 | 0.81 | 0.98 | 0.89 | 0.92 | 0.98 |
| HCK-78-138 | 0.87 | 0.98 | 0.91 | 0.95 | 0.98 |
| INTERSECTIN1-1002-1060 | 0.86 | 0.98 | 0.86 | 0.92 | 0.97 |
| INTERSECTIN1-1074-1138 | 0.61 | 0.99 | 0.73 | 0.72 | 0.97 |
| INTERSECTIN1-1155-1214 | 0.45 | 1 | 0.78 | 0.62 | 0.89 |
| INTERSECTIN1-740-806 | 0.69 | 0.99 | 0.94 | 0.9 | 0.97 |
| INTERSECTIN1-913-971 | 0.09 | 1 | 0.38 | 0.23 | 0.91 |
| IRSP53-374-437 | 0.67 | 0.95 | 0.69 | 0.75 | 0.93 |
| LCK-61-121 | 0.55 | 1 | 0.86 | 0.72 | 0.94 |
| LYN-63-123 | 0.92 | 0.92 | 0.91 | 0.98 | 0.98 |

*Continued on next page .....*

**Table A.2.3** – *Continued from previous page*

| SH3 domain | Sensitivity | Specificity | Precision | AUC PR | AUC ROC |
|---|---|---|---|---|---|
| MLK3-41-105 | 0.5 | 1 | 0.79 | 0.68 | 0.96 |
| MPP1-158-228 | 0.52 | 0.99 | 0.63 | 0.64 | 0.97 |
| MYO7A-1603-1672 | 0.56 | 0.99 | 0.67 | 0.65 | 0.97 |
| NCF1-226-285 | 0.47 | 1 | 0.75 | 0.67 | 0.98 |
| NCK2-195-257 | 0.56 | 0.99 | 0.73 | 0.72 | 0.97 |
| NPHP1-152-212 | 0.83 | 0.99 | 0.9 | 0.92 | 0.98 |
| N-SRC-84-145 | 0.64 | 0.99 | 0.69 | 0.72 | 0.97 |
| OSTF1-12-71 | 0.83 | 0.97 | 0.89 | 0.93 | 0.97 |
| P51NOX-399-458 | 0.2 | 1 | 0.54 | 0.45 | 0.96 |
| PAC2-426-486 | 0.83 | 0.98 | 0.9 | 0.95 | 0.98 |
| PAC3-363-424 | 0.73 | 0.97 | 0.86 | 0.89 | 0.96 |
| PIK3R1-3-79 | 0.86 | 0.98 | 0.91 | 0.95 | 0.98 |
| PLCG1-791-851 | 0.79 | 0.97 | 0.89 | 0.93 | 0.97 |
| RASGAP-279-341 | 0.8 | 0.98 | 0.81 | 0.89 | 0.98 |
| RIMB1-1625-1693 | 0.74 | 1 | 0.9 | 0.87 | 0.97 |
| RIMB1-1764-1831 | 0.69 | 1 | 0.98 | 0.82 | 0.92 |
| RUSC1-844-902 | 0.1 | 1 | 0.45 | 0.29 | 0.95 |
| SH3PX3-1-61 | 0.83 | 0.98 | 0.91 | 0.95 | 0.98 |
| SNX18-1-61 | 0.79 | 0.99 | 0.93 | 0.93 | 0.97 |
| SNX9-1-62 | 0.86 | 0.97 | 0.89 | 0.93 | 0.97 |
| SRC-84-145 | 0.91 | 0.93 | 0.85 | 0.96 | 0.98 |
| STAM1-210-269 | 0.84 | 0.99 | 0.83 | 0.91 | 0.98 |
| STAM2-202-261 | 0.1 | 1 | 0.2 | 0.26 | 0.84 |
| TUBA-145-204 | 0.7 | 0.98 | 0.84 | 0.86 | 0.96 |
| TUBA-1513-1576 | 0.65 | 0.98 | 0.84 | 0.81 | 0.93 |
| TUBA-2-61 | 0.36 | 1 | 0.65 | 0.53 | 0.9 |

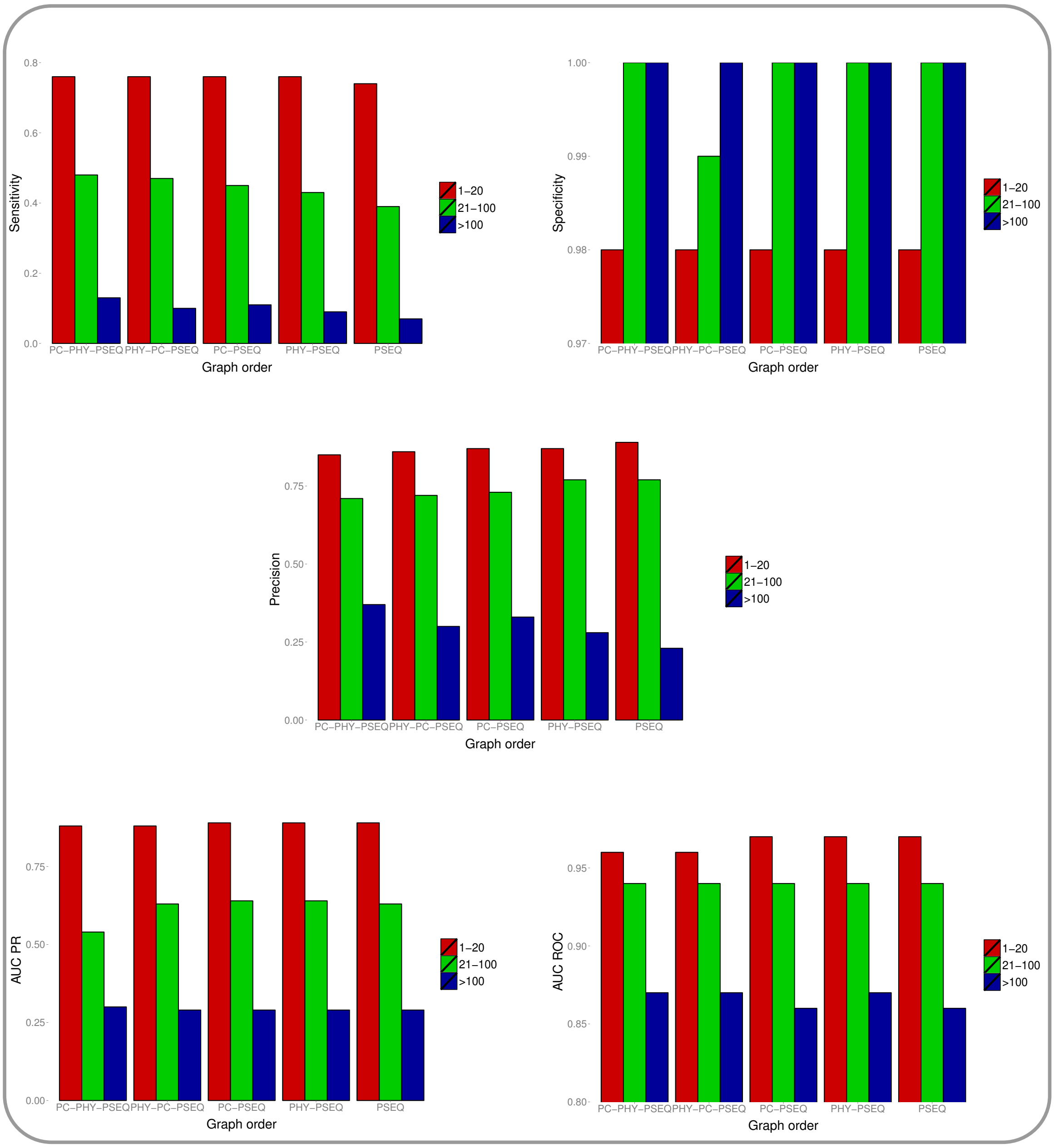


**Figure A.2.3.:** Here, we compare the performances (sensitivity, specificity, precision, AUC PR, and AUC ROC) using different encoding orders, i.e., PC-PHY-PSEQ, PHY-PC-PSEQ, PC-PSEQ, PHY-PSEQ, and PSEQ (PC = peptide charges, PHY = peptide hydrophobicity, and PSEQ = peptide sequence). Red bars indicate the performance of the domains having the negative/positive ratio 1 to 20, green bars indicate the performance of the domains having the negative/positive ratio 21 to 100, and the blue bars indicate the performance of the domains having the negative/positive ratio more than 100. Overall, the graph indicates PC-PHY-PSEQ encoding order performs better. The figure is adapted from the supplementary materials of [P3].

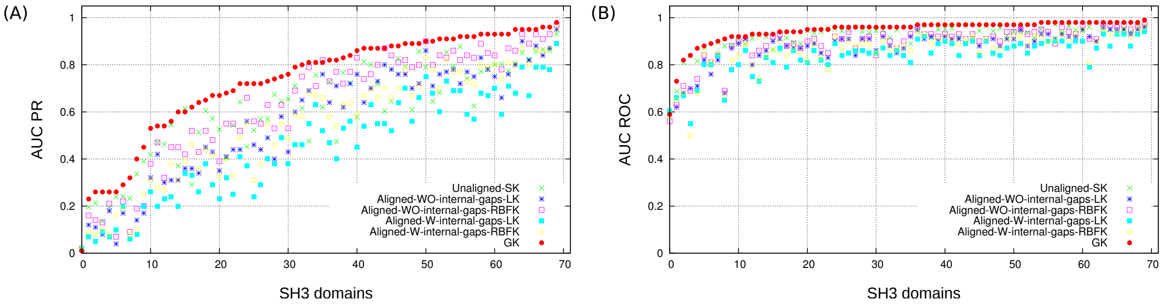


**Figure A.2.4.:** 10-fold cross-validation performance comparison with Graph Kernel (GK), unaligned sequences string kernel (Unaligned-SK), aligned sequences without internal gaps ("zero gaps") linear kernel (Aligned-WO-internal-gaps-LK), aligned sequences without internal gaps ("zero gaps") Gaussian kernel (Aligned-WO-internal-gaps-RBFK), aligned sequences with internal gaps linear kernel (Aligned-W-internal-gaps-LK), and aligned sequences with internal gaps Gaussian kernel (Aligned-W-internal-gaps-RBFK). The average AUC PR and AUC ROC are plotted for all 70 SH3 domains. The domains are sorted by increasing average performance for the Graph Kernel method. The figure is taken from the supplementary materials of [P3].

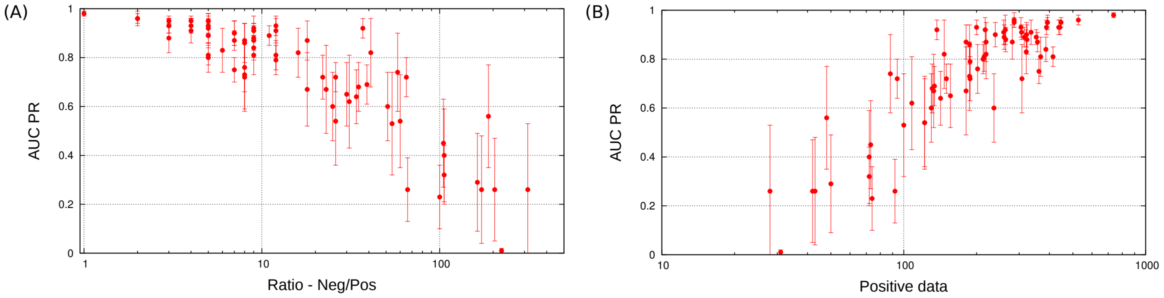


**Figure A.2.5.:** 10-fold cross-validation performance with filtered negative interactions for Single Domain Graph Kernel (GK). Area under the PR curve with standard deviation for Graph Kernel for each SH3 domain. The domains are sorted by increasing negative ratio (left) and positive interaction data (right). The error bars represent respective standard deviation. The figure is taken from the supplementary materials of [P3].

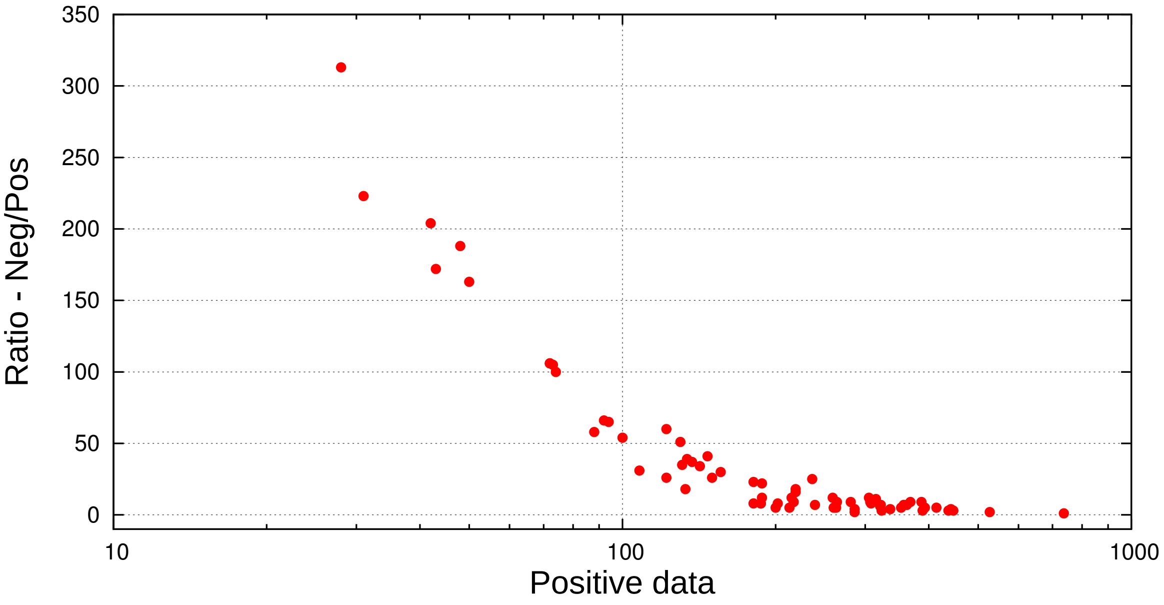


**Figure A.2.6.:** Here, we compare the number of positive data and the negative/positive ratio for each SH3 domain. The domains are sorted by increasing number of positive data. The figure indicates less imbalanced datasets (low negative/positive ratio) having higher number of positive data and also provide good performance. The figure is taken from the supplementary materials of [P3].

## A.3. PDZ domain data

**Figure A.3.1.:** (A) The AUC ROC and (B) the AUC PR curve obtained by sequence-based feature encoding (red line) and contact-based feature encoding (green dashed line) method. The figure is taken from the supplementary materials of [P2].

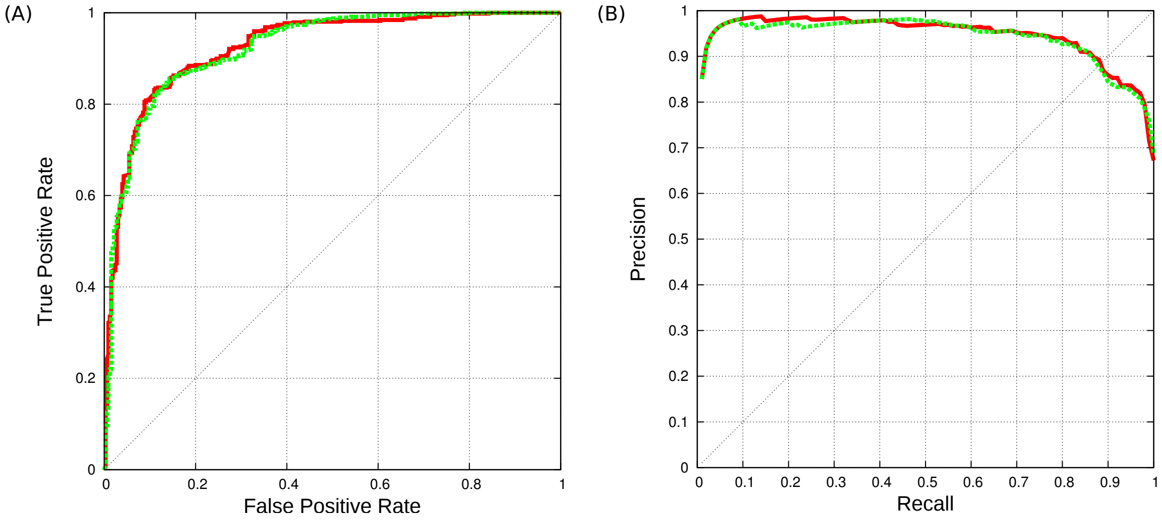


**Table A.3.1.:** Predictive performance of sequence-based approach. The table is taken from the supplementary materials of [P2].

| PDZ cluster | Positive int. | Sensitivity | Specificity | Precision | AUC PR | AUC ROC |
|---|---|---|---|---|---|---|
| 1 | 147 | 0.94 | 0.72 | 0.93 | 0.97 | 0.92 |
| 2 | 37 | 0.77 | 0.89 | 0.91 | 0.95 | 0.94 |
| 4 | 85 | 0.91 | 0.83 | 0.97 | 0.98 | 0.87 |
| 5 | 54 | 0.69 | 0.75 | 0.78 | 0.85 | 0.83 |
| 6 | 27 | 0.75 | 1 | 1 | 0.97 | 0.95 |
| 8 | 36 | 0.55 | 0.93 | 0.9 | 0.87 | 0.87 |
| 13 | 27 | 0.73 | 0.95 | 0.93 | 0.96 | 0.96 |
| 15 | 51 | 0.81 | 0.93 | 0.95 | 0.95 | 0.95 |
| 20 | 67 | 0.92 | 0.71 | 0.93 | 0.97 | 0.91 |
| 22 | 67 | 0.81 | 0.94 | 0.95 | 0.97 | 0.95 |
| 29 | 29 | 0.8 | 0.65 | 0.76 | 0.87 | 0.78 |
| 30 | 13 | 0.33 | 0.9 | 0.27 | 0.79 | 0.83 |
| 37 | 60 | 0.97 | 1 | 1 | 1 | 0.99 |
| 41 | 12 | 0.75 | 1 | 1 | 0.98 | 0.99 |
| 42 | 37 | 0.42 | 0.94 | 0.79 | 0.79 | 0.85 |
| 54 | 53 | 0.87 | 0.97 | 0.97 | 0.99 | 0.98 |
| 66 | 36 | 0.98 | 0.96 | 0.98 | 1 | 1 |
| 68 | 50 | 0.96 | 0.63 | 0.92 | 0.98 | 0.91 |
| 87 | 19 | 0.62 | 0.71 | 0.52 | 0.82 | 0.82 |
| 96 | 31 | 0.9 | 0.76 | 0.9 | 0.89 | 0.8 |
| 98 | 81 | 0.9 | 0.94 | 0.97 | 0.98 | 0.95 |
| 120 | 55 | 0.78 | 0.57 | 0.86 | 0.88 | 0.73 |

**Table A.3.2.:** Predictive performances of contact-based approach. The table is taken from the supplementary materials of [P2].

| PDZ cluster | Sensitivity | Specificity | Precision | AUC PR | AUC ROC |
|---|---|---|---|---|---|
| 1 | 0.94 | 0.79 | 0.95 | 0.97 | 0.92 |
| 2 | 0.76 | 0.88 | 0.88 | 0.95 | 0.94 |
| 20 | 0.92 | 0.68 | 0.92 | 0.98 | 0.94 |
| 42 | 0.53 | 0.9 | 0.78 | 0.73 | 0.79 |
| 54 | 0.88 | 0.98 | 0.98 | 0.99 | 0.99 |
| 120 | 0.84 | 0.44 | 0.84 | 0.9 | 0.77 |

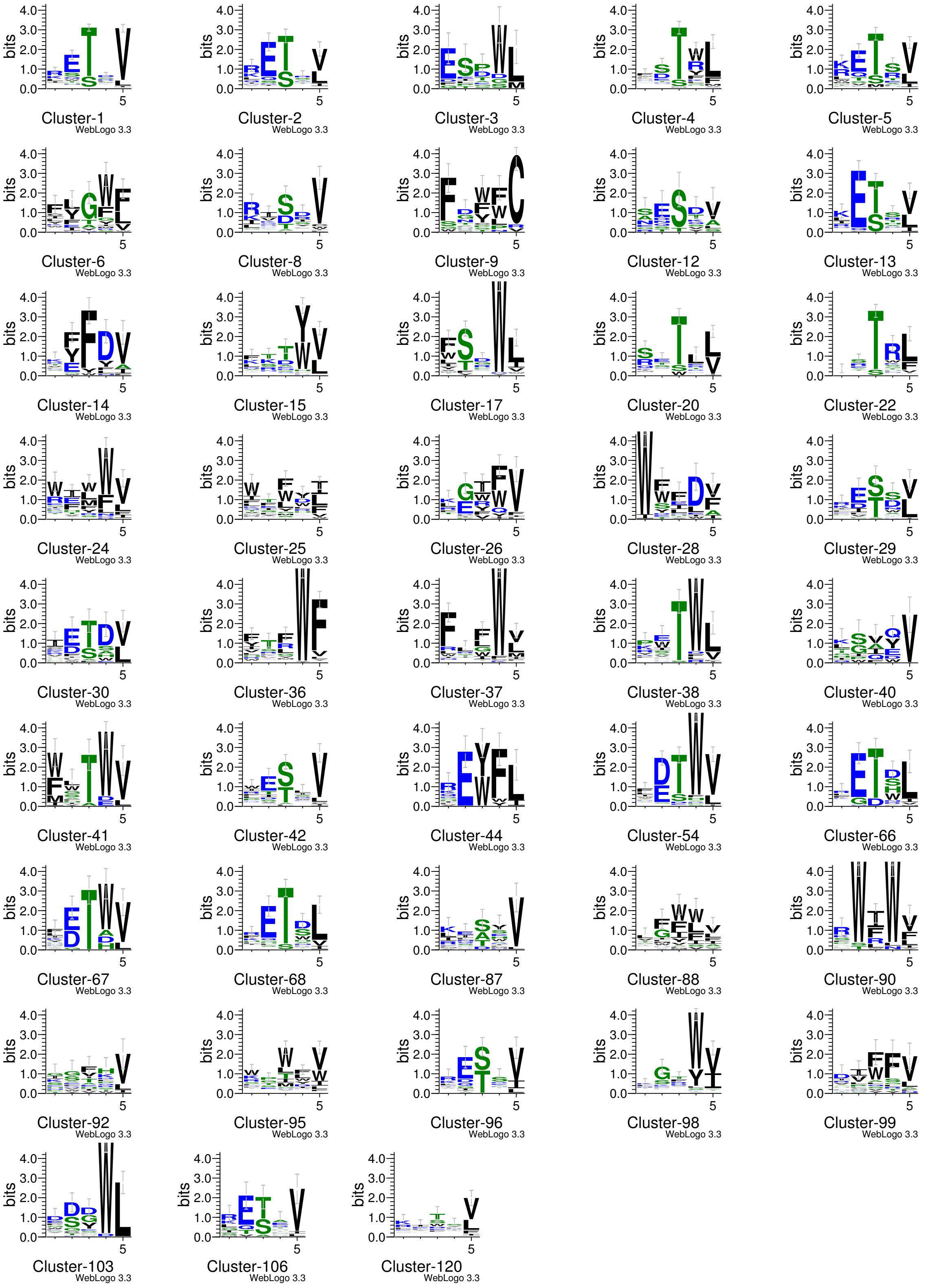
**Table A.3.3.:** Predicted binding peptides targeted by the highest number of PDZ domain in human. The table is taken from the supplementary materials of [P2].

| UniProt-ID | Peptide | Targeted by number of PDZ domains |
|---|---|---|
| Q8WXI2 | IETHV | 40 |
| Q14524 | RESIV | 39 |
| Q14957 | LESEV | 38 |
| Q9NYB5 | KETQL | 34 |
| P35354 | RSTEL | 34 |

**Table A.3.4.:** Predicted binding peptides targeted by the highest number of PDZ domain in mouse. The table is taken from the supplementary materials of [P2].

| UniProt-ID | Peptide | Targeted by number of PDZ domains |
|---|---|---|
| Q9ERB5 | KETRL | 42 |
| Q80YA9 | IETHV | 40 |
| O08911 | KETAL | 39 |
| Q01098 | LESEV | 38 |
| Q9JJV9 | RESIV | 37 |

**Figure A.3.2.:** Peptide logos for each PDZ domain family. `WebLogo` [180] was used for constructing the peptide logos. Different families show different ligand binding specificity. Families with at least 10 positive interactions were used for sequence logo. The figure is taken from the supplementary materials of [P2].



# Appendix B

## List of abbreviations

| | |
|---|---|
| AUC PR | Area under the curve precision and recall |
| AUC ROC | Area under the curve receiver operating characteristic |
| DAVID | Database for annotation, visualization and integrated discovery |
| EGFR | Epidermal growth factor receptor |
| EH | Eps15 homology |
| FN | False negative |
| FP | False positive |
| FPR | False positive rate |
| GIP | Graph isomorphism problem |
| GO | Gene Ontology |
| HTP | High-throughput |
| ITAM | Immunoreceptor tyrosine-based activation motif |
| IUP | Intrinsically unstructured protein |
| MCL | Markov clustering |
| MINT | Molecular interaction |
| ML | Machine learning |
| Pfam | Protein family |
| PID | Protein interaction domain |
| PRM | Peptide recognition module |
| PDB | Protein data bank |
| PDZ | PSD-95/DLG1/ZO-1 |
| PID | Protein interaction domain |
| PPI | Protein-protein interaction |
| PPII | Polyproline type II |
| PTB | Protein tyrosine binding |
| PTK | Protein tyrosine kinase |

| | |
|---|---|
| PTM | Post-translational modification |
| PTP | Protein tyrosine phosphatase |
| pTyr | Phosphotyrosine |
| PWM/PSSM | Position specific weight matrix/Position specific scoring matrix |
| RBF | Radial basis function |
| RTK | Receptor tyrosine kinase |
| SGD | Stochastic Gradient Descent |
| SH2 | Src homology 2 |
| SH3 | Src homology 3 |
| SSL | Semi-supervised learning |
| SVM | Support Vector Machine |
| TN | true negative |
| TP | true positive |
| TPR | true positive Rate |
| w.r.t. | with respect to |

# Standard amino acid abbreviations

| **Amino acid** | **Three-letter code** | **Single-letter code** |
|---|---|---|
| Alanine | Ala | A |
| Arginine | Arg | R |
| Asparagine | Asn | N |
| Aspartic acid | Asp | D |
| Cysteine | Cys | C |
| Glutamic acid | Glu | E |
| Glutamine | Gln | Q |
| Glycine | Gly | G |
| Histidine | His | H |
| Isoleucine | Ile | I |
| Leucine | Leu | L |
| Lysine | Lys | K |
| Methionine | Met | M |
| Phenylalanine | Phe | F |
| Proline | Pro | P |
| Serine | Ser | S |
| Threonine | Thr | T |
| Tryptophan | Trp | W |
| Tyrosine | Tyr | Y |
| Valine | Val | V |

# Bibliography

[1] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter, *Molecular Biology of the Cell, 4th edition.* Garland Scince, 2002.

[2] N. E. Hynes, P. W. Ingham, W. A. Lim, C. J. Marshall, J. Massague, and T. Pawson, "Signalling change: signal transduction through the decades," *Nat Rev Mol Cell Biol*, vol. 14, no. 6, pp. 393–8, 2013.

[3] I. Sadowski, J. C. Stone, and T. Pawson, "A noncatalytic domain conserved among cytoplasmic protein-tyrosine kinases modifies the kinase function and transforming activity of Fujinami sarcoma virus P130gag-fps," *Mol Cell Biol*, vol. 6, no. 12, pp. 4396–408, 1986.

[4] E. B. Haura, "From modules to medicine: How modular domains and their associated networks can enable personalized medicine," *FEBS Lett*, vol. 586, no. 17, pp. 2580–5, 2012.

[5] J. Ciesla, T. Fraczyk, and W. Rode, "Phosphorylation of basic amino acid residues in proteins: important but easily missed," *Acta Biochim Pol*, vol. 58, no. 2, pp. 137–48, 2011.

[6] J. Schlessinger and M. A. Lemmon, "SH2 and PTB domains in tyrosine kinase signaling," *Sci STKE*, vol. 2003, no. 191, p. RE12, 2003.

[7] A. C. Porter and R. R. Vaillancourt, "Tyrosine kinase receptor-activated signal transduction pathways which lead to oncogenesis," *Oncogene*, vol. 17, no. 11 Reviews, pp. 1343–52, 1998.

[8] Y. Yarden and M. X. Sliwkowski, "Untangling the ErbB signalling network," *Nat Rev Mol Cell Biol*, vol. 2, no. 2, pp. 127–37, 2001.

[9] P. Blume-Jensen and T. Hunter, "Oncogenic kinase signalling," *Nature*, vol. 411, no. 6835, pp. 355–65, 2001.

[10] B. T. Seet, I. Dikic, M.-M. Zhou, and T. Pawson, "Reading protein modifications with interaction domains," *Nat Rev Mol Cell Biol*, vol. 7, no. 7, pp. 473–83, 2006.

[11] T. Pawson and J. D. Scott, "Signaling through scaffold, anchoring, and adaptor proteins," *Science*, vol. 278, no. 5346, pp. 2075–80, 1997.

[12] T. Pawson and G. D. Gish, "SH2 and SH3 domains: from structure to function," *Cell*, vol. 71, no. 3, pp. 359–62, 1992.

[13] G. W. Booker, A. L. Breeze, A. K. Downing, G. Panayotou, I. Gout, M. D. Waterfield, and I. D. Campbell, "Structure of an SH2 domain of the p85 alpha subunit of phosphatidylinositol-3-OH kinase," *Nature*, vol. 358, no. 6388, pp. 684–7, 1992.

[14] M. Overduin, C. B. Rios, B. J. Mayer, D. Baltimore, and D. Cowburn, "Three-dimensional solution structure of the src homology 2 domain of c-abl," *Cell*, vol. 70, no. 4, pp. 697–704, 1992.

[15] B. J. Mayer, M. Hamaguchi, and H. Hanafusa, "A novel viral oncogene with structural similarity to phospholipase C," *Nature*, vol. 332, no. 6161, pp. 272–5, 1988.

[16] D. Anderson, C. A. Koch, L. Grey, C. Ellis, M. F. Moran, and T. Pawson, "Binding of SH2 domains of phospholipase C gamma 1, GAP, and Src to activated growth factor receptors," *Science*, vol. 250, no. 4983, pp. 979–82, 1990.

[17] W. A. Lim and T. Pawson, "Phosphotyrosine signaling: evolving a new cellular communication system," *Cell*, vol. 142, no. 5, pp. 661–7, 2010.

[18] B. A. Liu, E. Shah, K. Jablonowski, A. Stergachis, B. Engelmann, and P. D. Nash, "The SH2 domain-containing proteins in 21 species establish the provenance and scope of phosphotyrosine signaling in eukaryotes," *Sci Signal*, vol. 4, no. 202, p. ra83, 2011.

[19] M. Magrane and U. Consortium, "UniProt Knowledgebase: a hub of integrated protein data," *Database (Oxford)*, vol. 2011, p. bar009, 2011.

[20] B. A. Liu, K. Jablonowski, M. Raina, M. Arce, T. Pawson, and P. D. Nash, "The human and mouse complement of SH2 domain proteins-establishing the boundaries of phosphotyrosine signaling," *Mol Cell*, vol. 22, no. 6, pp. 851–68, 2006.

[21] T. Pawson, "Specificity in signal transduction: from phosphotyrosine-SH2 domain interactions to complex cellular systems," *Cell*, vol. 116, no. 2, pp. 191–203, 2004.

[22] B. J. Mayer, P. K. Jackson, and D. Baltimore, "The noncatalytic src homology region 2 segment of abl tyrosine kinase binds to tyrosine-phosphorylated cellular proteins with high affinity," *Proc Natl Acad Sci USA*, vol. 88, no. 2, pp. 627–31, 1991.

[23] Z. Songyang, S. E. Shoelson, M. Chaudhuri, G. Gish, T. Pawson, W. G. Haser, F. King, T. Roberts, S. Ratnofsky, and R. J. Lechleider, "SH2 domains recognize specific phosphopeptide sequences," *Cell*, vol. 72, no. 5, pp. 767–78, 1993.

[24] Z. Songyang and L. C. Cantley, "Recognition and specificity in protein tyrosine kinase-mediated signalling," *Trends in Biochemical Sciences*, vol. 20, no. 11, pp. 470–5, 1995.

[25] L. C. Cantley and Z. Songyang, "Specificity in recognition of phosphopeptides by src-homology 2 domains.," *J Cell Sci Suppl*, vol. 18, pp. 121–126, 1994.

[26] F. Poy, M. B. Yaffe, J. Sayos, K. Saxena, M. Morra, J. Sumegi, L. C. Cantley, C. Terhorst, and M. J. Eck, "Crystal structures of the XLP protein SAP reveal a class of SH2 domains with extended, phosphotyrosine-independent sequence recognition," *Mol Cell*, vol. 4, no. 4, pp. 555–61, 1999.

[27] J. Sayos, C. Wu, M. Morra, N. Wang, X. Zhang, D. Allen, S. van Schaik, L. Notarangelo, R. Geha, M. G. Roncarolo, H. Oettgen, J. E. De Vries, G. Aversa, and C. Terhorst, "The X-linked lymphoproliferative-disease gene product SAP regulates signals induced through the co-receptor SLAM," *Nature*, vol. 395, no. 6701, pp. 462–9, 1998.

[28] M. Tartaglia, E. L. Mehler, R. Goldberg, G. Zampino, H. G. Brunner, H. Kremer, I. van der Burgt, A. H. Crosby, A. Ion, S. Jeffery, K. Kalidas, M. A. Patton, R. S. Kucherlapati, and B. D. Gelb, "Mutations in PTPN11, encoding the protein tyrosine phosphatase SHP-2, cause Noonan syndrome," *Nat Genet*, vol. 29, no. 4, pp. 465–8, 2001.

[29] S. R. Tzeng, M. T. Pai, F. D. Lung, C. W. Wu, P. P. Roller, B. Lei, C. J. Wei, S. C. Tu, S. H. Chen, W. J. Soong, and J. W. Cheng, "Stability and peptide binding specificity of Btk SH2 domain: molecular basis for X-linked agammaglobulinemia," *Protein Sci*, vol. 9, no. 12, pp. 2377–85, 2000.

[30] E. Friedman, P. V. Gejman, G. A. Martin, and F. McCormick, "Nonsense mutations in the C-terminal SH2 region of the GTPase activating protein (GAP) gene in human tumours," *Nat Genet*, vol. 5, no. 3, pp. 242–7, 1993.

[31] M. A. Larkin, G. Blackshields, N. P. Brown, R. Chenna, P. A. McGettigan, H. McWilliam, F. Valentin, I. M. Wallace, A. Wilm, R. Lopez, J. D. Thompson, T. J. Gibson, and D. G. Higgins, "Clustal W and Clustal X version 2.0," *Bioinformatics*, vol. 23, no. 21, pp. 2947–8, 2007.

[32] I. Letunic and P. Bork, "Interactive Tree Of Life v2: online annotation and display of phylogenetic trees made easy," *Nucleic Acids Res*, vol. 39, no. Web Server issue, pp. W475–8, 2011.

[33] G. Waksman, D. Kominos, S. C. Robertson, N. Pant, D. Baltimore, R. B. Birge, D. Cowburn, H. Hanafusa, B. J. Mayer, M. Overduin, M. D. Resh, C. B. Rios, L. Silverman, and J. Kuriyan, "Crystal structure of the phosphotyrosine recognition domain SH2 of v-src complexed with tyrosine-phosphorylated peptides," *Nature*, vol. 358, no. 6388, pp. 646–53, 1992.

[34] G. Waksman, S. E. Shoelson, N. Pant, D. Cowburn, and J. Kuriyan, "Binding of a high affinity phosphotyrosyl peptide to the Src SH2 domain: crystal structures of the complexed and peptide-free forms," *Cell*, vol. 72, no. 5, pp. 779–90, 1993.

[35] T. Kaneko, R. Joshi, S. M. Feller, and S. S. Li, "Phosphotyrosine recognition domains: the typical, the atypical and the versatile," *Cell Commun Signal*, vol. 10, no. 1, p. 32, 2012.

[36] B. A. Liu, B. W. Engelmann, and P. D. Nash, "The language of SH2 domain interactions defines phosphotyrosine-mediated signal transduction," *FEBS Lett*, 2012.

[37] T. Kaneko, H. Huang, B. Zhao, L. Li, H. Liu, C. K. Voss, C. Wu, M. R. Schiller, and S. S.-C. Li, "Loops govern SH2 domain specificity by controlling access to binding pockets," *Sci Signal*, vol. 3, no. 120, p. ra34, 2010.

[38] K. B. Bibbins, H. Boeuf, and H. E. Varmus, "Binding of the Src SH2 domain to phosphopeptides is determined by residues in both the SH2 domain and the phosphopeptides," *Mol Cell Biol*, vol. 13, no. 12, pp. 7278–87, 1993.

[39] H. Huang, L. Li, C. Wu, D. Schibli, K. Colwill, S. Ma, C. Li, P. Roy, K. Ho, Z. Songyang, T. Pawson, Y. Gao, and S. S.-C. Li, "Defining the specificity space of the human SRC homology 2 domain," *Mol Cell Proteomics*, vol. 7, no. 4, pp. 768–84, 2008.

[40] D. Imhof, A.-S. Wavreille, A. May, M. Zacharias, S. Tridandapani, and D. Pei, "Sequence specificity of SHP-1 and SHP-2 Src homology 2 domains. Critical roles of residues beyond the pY+3 position," *Journal of Biological Chemistry*, vol. 281, no. 29, pp. 20271–82, 2006.

[41] T. Kaneko, H. Huang, X. Cao, X. Li, C. Li, C. Voss, S. S. Sidhu, and S. S. C. Li, "Superbinder SH2 domains act as antagonists of cell signaling," *Sci Signal*, vol. 5, no. 243, p. ra68, 2012.

[42] R. C. Edgar, "MUSCLE: multiple sequence alignment with high accuracy and high throughput," *Nucleic Acids Res*, vol. 32, no. 5, pp. 1792–7, 2004.

[43] B. A. Liu, K. Jablonowski, E. E. Shah, B. W. Engelmann, R. B. Jones, and P. D. Nash, "SH2 domains recognize contextual peptide sequence information to determine selectivity," *Mol Cell Proteomics*, vol. 9, no. 11, pp. 2391–404, 2010.

[44] A. Kaushansky, A. Gordus, B. Chang, J. Rush, and G. MacBeath, "A quantitative study of the recruitment potential of all intracellular tyrosine residues on EGFR, FGFR1 and IGF1R," *Mol Biosyst*, vol. 4, no. 6, pp. 643–53, 2008.

[45] A. Charest, J. Wagner, S. Jacob, C. J. McGlade, and M. L. Tremblay, "Phosphotyrosine-independent binding of SHC to the NPLH sequence of murine protein-tyrosine phosphatase-PEST. Evidence for extended phosphotyrosine binding/phosphotyrosine interaction domain recognition specificity," *Journal of Biological Chemistry*, vol. 271, no. 14, pp. 8424–9, 1996.

[46] K. Dai, S. Liao, J. Zhang, X. Zhang, and X. Tu, "Solution structure of tensin2 SH2 domain and its phosphotyrosine-independent interaction with DLC-1," *PLoS One*, vol. 6, no. 7, p. e21965, 2011.

[47] Y.-C. Liao, L. Si, R. W. deVere White, and S. H. Lo, "The phosphotyrosine-independent interaction of DLC-1 and the SH2 domain of cten regulates focal adhesion localization and growth suppression activity of DLC-1," *J Cell Biol*, vol. 176, no. 1, pp. 43–9, 2007.

[48] E. F. Pettersen, T. D. Goddard, C. C. Huang, G. S. Couch, D. M. Greenblatt, E. C. Meng, and T. E. Ferrin, "UCSF Chimera–a visualization system for exploratory research and analysis," *J Comput Chem*, vol. 25, no. 13, pp. 1605–12, 2004.

[49] B. J. Mayer, "SH3 domains: complexity in moderation," *J Cell Sci*, vol. 114, no. Pt 7, pp. 1253–63, 2001.

[50] A. Musacchio, M. Noble, R. Pauptit, R. Wierenga, and M. Saraste, "Crystal structure of a Src-homology 3 (SH3) domain," *Nature*, vol. 359, no. 6398, pp. 851–5, 1992.

[51] W. A. Lim, F. M. Richards, and R. O. Fox, "Structural determinants of peptide-binding orientation and of sequence specificity in SH3 domains," *Nature*, vol. 372, no. 6504, pp. 375–9, 1994.

[52] M. L. Stahl, C. R. Ferenz, K. L. Kelleher, R. W. Kriz, and J. L. Knopf, “Sequence similarity of phospholipase C with the non-catalytic region of src,” *Nature*, vol. 332, no. 6161, pp. 269–72, 1988.

[53] V. P. Lehto, V. M. Wasenius, P. Salven, and M. Saraste, “Transforming and membrane proteins,” *Nature*, vol. 334, no. 6181, p. 388, 1988.

[54] S. Karkkainen, M. Hiipakka, J.-H. Wang, I. Kleino, M. Vaha-Jaakkola, G. H. Renkema, M. Liss, R. Wagner, and K. Saksela, “Identification of preferred protein interactions by phage-display of the human Src homology-3 proteome,” *EMBO Rep*, vol. 7, no. 2, pp. 186–91, 2006.

[55] M. Hiipakka, K. Poikonen, and K. Saksela, “SH3 domains with high affinity and engineered ligand specificities targeted to HIV-1 Nef,” *J Mol Biol*, vol. 293, no. 5, pp. 1097–106, 1999.

[56] T. Kaneko, L. Li, and S. S.-C. Li, “The SH3 domain–a family of versatile peptide- and protein-recognition module,” *Front Biosci*, vol. 13, pp. 4938–52, 2008.

[57] B. K. Kay, “SH3 domains come of age,” *FEBS Lett*, vol. 586, no. 17, pp. 2606–8, 2012.

[58] M. Carducci, L. Perfetto, L. Briganti, S. Paoluzi, S. Costa, J. Zerweck, M. Schutkowski, L. Castagnoli, and G. Cesareni, “The protein interaction network mediated by human SH3 domains,” *Biotechnol Adv*, vol. 30, no. 1, pp. 4–15, 2012.

[59] G. Cesareni, S. Panni, G. Nardelli, and L. Castagnoli, “Can we infer peptide recognition specificity mediated by SH3 domains?,” *FEBS Lett*, vol. 513, no. 1, pp. 38–44, 2002.

[60] D. Grabulovski, M. Kaspar, and D. Neri, “A novel, non-immunogenic Fyn SH3-derived binding protein with tumor vascular targeting properties,” *Journal of Biological Chemistry*, vol. 282, no. 5, pp. 3196–204, 2007.

[61] M. Hiipakka and K. Saksela, “Versatile retargeting of SH3 domain binding by modification of non-conserved loop residues,” *FEBS Lett*, vol. 581, no. 9, pp. 1735–41, 2007.

[62] S. S.-C. Li, “Specificity and versatility of SH3 and other proline-recognition domains: structural basis and implications for cellular signal transduction,” *Biochem J*, vol. 390, no. Pt 3, pp. 641–53, 2005.

[63] H. Yu, J. K. Chen, S. Feng, D. C. Dalgarno, A. W. Brauer, and S. L. Schreiber, “Structural basis for the binding of proline-rich peptides to SH3 domains,” *Cell*, vol. 76, no. 5, pp. 933–45, 1994.

[64] S. Feng, J. K. Chen, H. Yu, J. A. Simon, and S. L. Schreiber, “Two binding orientations for peptides to the Src SH3 domain: development of a general model for SH3-ligand interactions,” *Science*, vol. 266, no. 5188, pp. 1241–7, 1994.

[65] S. Feng, C. Kasahara, R. J. Rickles, and S. L. Schreiber, “Specific interactions outside the proline-rich core of two classes of Src homology 3 ligands,” *Proc Natl Acad Sci USA*, vol. 92, no. 26, pp. 12408–15, 1995.

[66] T. Kesti, A. Ruppelt, J.-H. Wang, M. Liss, R. Wagner, K. Tasken, and K. Saksela, “Reciprocal regulation of SH3 and SH2 domain binding via tyrosine phosphorylation of a common site in CD3epsilon,” *J Immunol*, vol. 179, no. 2, pp. 878–85, 2007.

[67] A. M. Mongiovi, P. R. Romano, S. Panni, M. Mendoza, W. T. Wong, A. Musacchio, G. Cesareni, and P. P. Di Fiore, “A novel peptide-SH3 interaction,” *EMBO J*, vol. 18, no. 19, pp. 5300–9, 1999.

[68] B. Matoskova, W. T. Wong, A. E. Salcini, P. G. Pelicci, and P. P. Di Fiore, “Constitutive phosphorylation of eps8 in tumor cell lines: relevance to malignant transformation,” *Mol Cell Biol*, vol. 15, no. 7, pp. 3805–12, 1995.

[69] F. Fazioli, L. Minichiello, V. Matoska, P. Castagnino, T. Miki, W. T. Wong, and P. P. Di Fiore, “Eps8, a substrate for the epidermal growth factor receptor kinase, enhances EGF-dependent mitogenic signals,” *EMBO J*, vol. 12, no. 10, pp. 3799–808, 1993.

[70] R. J. Rickles, M. C. Botfield, Z. Weng, J. A. Taylor, O. M. Green, J. S. Brugge, and M. J. Zoller, “Identification of Src, Fyn, Lyn, PI3K and Abl SH3 domain ligands using phage display libraries,” *EMBO J*, vol. 13, no. 23, pp. 5598–604, 1994.

[71] M. Kato, K. Miyazawa, and N. Kitamura, “A deubiquitinating enzyme UBPY interacts with the Src homology 3 domain of Hrs-binding protein via a novel binding motif PX(V/I)(D/N)RXXKP,” *Journal of Biological Chemistry*, vol. 275, no. 48, pp. 37481–7, 2000.

[72] Q. Liu, D. Berry, P. Nash, T. Pawson, C. J. McGlade, and S. S.-C. Li, “Structural basis for specific binding of the Gads SH3 domain to an RxxK motif-containing SLP-76 peptide: a novel mode of peptide recognition,” *Mol Cell*, vol. 11, no. 2, pp. 471–81, 2003.

[73] B. T. Seet, D. M. Berry, J. S. Maltzman, J. Shabason, M. Raina, G. A. Koretzky, C. J. McGlade, and T. Pawson, “Efficient T-cell receptor signaling requires a high-affinity interaction between the Gads C-SH3 domain and the SLP-76 RxxK motif,” *EMBO J*, vol. 26, no. 3, pp. 678–89, 2007.

[74] L. Tian, L. Chen, H. McClafferty, C. A. Sailer, P. Ruth, H.-G. Knaus, and M. J. Shipston, “A noncanonical SH3 domain binding motif links BK channels to the actin cytoskeleton via the SH3 adapter cortactin,” *FASEB J*, vol. 20, no. 14, pp. 2588–90, 2006.

[75] G. Moncalian, N. Cardenes, Y. L. Deribe, M. Spinola-Amilibia, I. Dikic, and J. Bravo, “Atypical polyproline recognition by the CMS N-terminal Src homology 3 domain,” *Journal of Biological Chemistry*, vol. 281, no. 50, pp. 38845–53, 2006.

[76] H. Kang, C. Freund, J. S. Duke-Cohan, A. Musacchio, G. Wagner, and C. E. Rudd, “SH3 domain recognition of a proline-independent tyrosine-based RKxxYxxY motif in immune cell adaptor SKAP55,” *EMBO J*, vol. 19, no. 12, pp. 2889–99, 2000.

[77] A. H. Y. Tong, B. Drees, G. Nardelli, G. D. Bader, B. Brannetti, L. Castagnoli, M. Evangelista, S. Ferracuti, B. Nelson, S. Paoluzi, M. Quondam, A. Zucconi, C. W. V. Hogue, S. Fields, C. Boone, and G. Cesareni, “A combined experimental and computational strategy to define protein interaction networks for peptide recognition modules,” *Science*, vol. 295, no. 5553, pp. 321–4, 2002. Paper as Print Copy.

[78] A. S. Shaw and E. L. Filbert, “Scaffold proteins and immune-cell signalling,” *Nat Rev Immunol*, vol. 9, no. 1, pp. 47–56, 2009.

[79] M. C. Good, J. G. Zalatan, and W. A. Lim, "Scaffold proteins: hubs for controlling the flow of cellular information," *Science*, vol. 332, no. 6030, pp. 680–6, 2011.

[80] T. Pawson and P. Nash, "Assembly of cell regulatory systems through protein interaction domains," *Science*, vol. 300, no. 5618, pp. 445–52, 2003.

[81] C. P. Ponting, "Evidence for PDZ domains in bacteria, yeast, and plants," *Protein Sci*, vol. 6, no. 2, pp. 464–8, 1997.

[82] C. Nourry, S. G. N. Grant, and J.-P. Borg, "PDZ domain proteins: plug and play!," *Sci STKE*, vol. 2003, no. 179, p. RE7, 2003.

[83] K. K. Dev, "Making protein interactions druggable: targeting PDZ domains," *Nat Rev Drug Discov*, vol. 3, no. 12, pp. 1047–56, 2004.

[84] E. Kim, M. Niethammer, A. Rothschild, Y. N. Jan, and M. Sheng, "Clustering of Shaker-type K+ channels by interaction with a family of membrane-associated guanylate kinases," *Nature*, vol. 378, no. 6552, pp. 85–8, 1995.

[85] K. O. Cho, C. A. Hunt, and M. B. Kennedy, "The rat brain postsynaptic density fraction contains a homolog of the Drosophila discs-large tumor suppressor protein," *Neuron*, vol. 9, no. 5, pp. 929–42, 1992.

[86] D. F. Woods and P. J. Bryant, "ZO-1, DlgA and PSD-95/SAP90: homologous proteins in tight, septate and synaptic cell junctions," *Mech Dev*, vol. 44, no. 2-3, pp. 85–9, 1993.

[87] M. B. Kennedy, "Origin of PDZ (DHR, GLGF) domains," *Trends in Biochemical Sciences*, vol. 20, no. 9, p. 350, 1995.

[88] J. R. Chen, B. H. Chang, J. E. Allen, M. A. Stiffler, and G. MacBeath, "Predicting PDZ domain-peptide interactions from primary sequences," *Nat Biotechnol*, vol. 26, no. 9, pp. 1041–5, 2008.

[89] Y. Ivarsson, "Plasticity of PDZ domains in ligand recognition and signaling," *FEBS Lett*, vol. 586, no. 17, pp. 2638–47, 2012.

[90] J. H. Morais Cabral, C. Petosa, M. J. Sutcliffe, S. Raza, O. Byron, F. Poy, S. M. Marfatia, A. H. Chishti, and R. C. Liddington, "Crystal structure of a PDZ domain," *Nature*, vol. 382, no. 6592, pp. 649–52, 1996.

[91] A. Ernst, S. L. Sazinsky, S. Hui, B. Currell, M. Dharsee, S. Seshagiri, G. D. Bader, and S. S. Sidhu, "Rapid evolution of functional complexity in a domain family," *Sci Signal*, vol. 2, no. 87, p. ra50, 2009.

[92] Z. Songyang, A. S. Fanning, C. Fu, J. Xu, S. M. Marfatia, A. H. Chishti, A. Crompton, A. C. Chan, J. M. Anderson, and L. C. Cantley, "Recognition of unique carboxyl-terminal motifs by distinct PDZ domains," *Science*, vol. 275, no. 5296, pp. 73–7, 1997.

[93] D. A. Doyle, A. Lee, J. Lewis, E. Kim, M. Sheng, and R. MacKinnon, "Crystal structures of a complexed and peptide-free membrane protein-binding domain: molecular basis of peptide recognition by PDZ," *Cell*, vol. 85, no. 7, pp. 1067–76, 1996.

[94] D. L. Daniels, A. R. Cohen, J. M. Anderson, and A. T. Brunger, “Crystal structure of the hCASK PDZ domain reveals the structural basis of class II PDZ domain target recognition,” *Nat Struct Biol*, vol. 5, no. 4, pp. 317–25, 1998.

[95] J. Grembecka, T. Cierpicki, Y. Devedjiev, U. Derewenda, B. S. Kang, J. H. Bushweller, and Z. S. Derewenda, “The binding of the PDZ tandem of syntenin to target proteins,” *Biochemistry*, vol. 45, no. 11, pp. 3674–83, 2006.

[96] T. Sugi, T. Oyama, T. Muto, S. Nakanishi, K. Morikawa, and H. Jingami, “Crystal structures of autoinhibitory PDZ domain of Tamalin: implications for metabotropic glutamate receptor trafficking regulation,” *EMBO J*, vol. 26, no. 8, pp. 2192–205, 2007.

[97] R. Tonikian, Y. Zhang, S. L. Sazinsky, B. Currell, J.-H. Yeh, B. Reva, H. A. Held, B. A. Appleton, M. Evangelista, Y. Wu, X. Xin, A. C. Chan, S. Seshagiri, L. A. Lasky, C. Sander, C. Boone, G. D. Bader, and S. S. Sidhu, “A specificity map for the PDZ domain family,” *PLoS Biol*, vol. 6, no. 9, p. e239, 2008.

[98] K. Luck, S. Fournane, B. Kieffer, M. Masson, Y. Nomine, and G. Trave, “Putting into practice domain-linear motif interaction predictions for exploration of protein networks,” *PLoS One*, vol. 6, no. 11, p. e25376, 2011.

[99] W. Feng, H. Wu, L.-N. Chan, and M. Zhang, “Par-3-mediated junctional localization of the lipid phosphatase PTEN is required for cell polarity establishment,” *Journal of Biological Chemistry*, vol. 283, no. 34, pp. 23440–9, 2008.

[100] N. L. Stricker, K. S. Christopherson, B. A. Yi, P. J. Schatz, R. W. Raab, G. Dawes, D. E. J. Bassett, D. S. Bredt, and M. Li, “PDZ domain of neuronal nitric oxide synthase recognizes novel C-terminal peptide sequences,” *Nat Biotechnol*, vol. 15, no. 4, pp. 336–42, 1997.

[101] H. C. Kornau, L. T. Schenker, M. B. Kennedy, and P. H. Seeburg, “Domain interaction between NMDA receptor subunits and the postsynaptic density protein PSD-95,” *Science*, vol. 269, no. 5231, pp. 1737–40, 1995.

[102] J. P. Borg, S. Marchetto, A. Le Bivic, V. Ollendorff, F. Jaulin-Bastard, H. Saito, E. Fournier, J. Adelaide, B. Margolis, and D. Birnbaum, “ERBIN: a basolateral PDZ protein that interacts with the mammalian ERBB2/HER2 receptor,” *Nat Cell Biol*, vol. 2, no. 7, pp. 407–14, 2000.

[103] J. J. Grootjans, P. Zimmermann, G. Reekmans, A. Smets, G. Degeest, J. Durr, and G. David, “Syntenin, a PDZ protein that binds syndecan cytoplasmic domains,” *Proc Natl Acad Sci USA*, vol. 94, no. 25, pp. 13683–8, 1997.

[104] B. J. Hillier, K. S. Christopherson, K. E. Prehoda, D. S. Bredt, and W. A. Lim, “Unexpected modes of PDZ domain scaffolding revealed by structure of nNOS-syntrophin complex,” *Science*, vol. 284, no. 5415, pp. 812–5, 1999.

[105] T. B. C. London, H.-J. Lee, Y. Shao, and J. Zheng, “Interaction between the internal motif KTXXXI of Idax and mDvl PDZ domain,” *Biochem Biophys Res Commun*, vol. 322, no. 1, pp. 326–32, 2004.

[106] X. Z. Xu, A. Choudhury, X. Li, and C. Montell, “Coordination of an array of signaling proteins through homo- and heteromeric interactions between PDZ domains and target proteins,” *J Cell Biol*, vol. 142, no. 2, pp. 545–55, 1998.

[107] A. G. Lau and R. A. Hall, “Oligomerization of NHERF-1 and NHERF-2 PDZ domains: differential regulation by association with receptor carboxyl-termini and by phosphorylation,” *Biochemistry*, vol. 40, no. 29, pp. 8572–80, 2001.

[108] P. Zimmermann, K. Meerschaert, G. Reekmans, I. Leenaerts, J. V. Small, J. Vandekerckhove, G. David, and J. Gettemans, “PIP(2)-PDZ domain binding controls the association of syntenin with the plasma membrane,” *Mol Cell*, vol. 9, no. 6, pp. 1215–25, 2002.

[109] H. Wu, W. Feng, J. Chen, L.-N. Chan, S. Huang, and M. Zhang, “PDZ domains of Par-3 as potential phosphoinositide signaling integrators,” *Mol Cell*, vol. 28, no. 5, pp. 886–98, 2007.

[110] D. R. Robinson, Y. M. Wu, and S. F. Lin, “The protein tyrosine kinase family of the human genome,” *Oncogene*, vol. 19, no. 49, pp. 5548–57, 2000.

[111] J. P. Montmayeur, M. Valius, J. Vandenheede, and A. Kazlauskas, “The platelet-derived growth factor beta receptor triggers multiple cytoplasmic signaling cascades that arrive at the nucleus as distinguishable inputs,” *Journal of Biological Chemistry*, vol. 272, no. 51, pp. 32670–8, 1997.

[112] M. B. Yaffe, “Phosphotyrosine-binding domains in signal transduction,” *Nat Rev Mol Cell Biol*, vol. 3, no. 3, pp. 177–86, 2002.

[113] E. J. Lowenstein, R. J. Daly, A. G. Batzer, W. Li, B. Margolis, R. Lammers, A. Ullrich, E. Y. Skolnik, D. Bar-Sagi, and J. Schlessinger, “The SH2 and SH3 domain-containing protein GRB2 links receptor tyrosine kinases to ras signaling,” *Cell*, vol. 70, no. 3, pp. 431–42, 1992.

[114] E. M. Bublil and Y. Yarden, “The EGF receptor family: spearheading a merger of signaling and therapeutics,” *Current Opinion in Cell Biology*, vol. 19, no. 2, pp. 124–34, 2007.

[115] C. Birchmeier and E. Gherardi, “Developmental roles of HGF/SF and its receptor, the c-Met tyrosine kinase,” *Trends Cell Biol*, vol. 8, no. 10, pp. 404–10, 1998.

[116] F. Maina, G. Pante, F. Helmbacher, R. Andres, A. Porthin, A. M. Davies, C. Ponzetto, and R. Klein, “Coupling Met to specific pathways results in distinct developmental outcomes,” *Mol Cell*, vol. 7, no. 6, pp. 1293–306, 2001.

[117] P. Hof, S. Pluskey, S. Dhe-Paganon, M. J. Eck, and S. E. Shoelson, “Crystal structure of the tyrosine phosphatase SHP-2,” *Cell*, vol. 92, no. 4, pp. 441–50, 1998.

[118] D. C. Saffran, O. Parolini, M. E. Fitch-Hilgenberg, D. J. Rawlings, D. E. Afar, O. N. Witte, and M. E. Conley, “Brief report: a point mutation in the SH2 domain of Bruton’s tyrosine kinase in atypical X-linked agammaglobulinemia,” *N Engl J Med*, vol. 330, no. 21, pp. 1488–91, 1994.

[119] M. Vidal, V. Gigoux, and C. Garbay, “SH2 and SH3 domains as targets for anti-proliferative agents,” *Crit Rev Oncol Hematol*, vol. 40, no. 2, pp. 175–86, 2001.

[120] S. L. Nix, A. H. Chishti, J. M. Anderson, and Z. Walther, “hCASK and hDlg associate in epithelia, and their src homology 3 and guanylate kinase domains participate in both intramolecular and intermolecular interactions,” *Journal of Biological Chemistry*, vol. 275, no. 52, pp. 41192–200, 2000.

[121] D. F. Woods, C. Hough, D. Peel, G. Callaini, and P. J. Bryant, “Dlg protein is required for junction structure, cell polarity, and proliferation control in Drosophila epithelia,” *J Cell Biol*, vol. 134, no. 6, pp. 1469–82, 1996.

[122] B. L. Anderson, I. Boldogh, M. Evangelista, C. Boone, L. A. Greene, and L. A. Pon, “The Src homology domain 3 (SH3) of a yeast type I myosin, Myo5p, binds to verprolin and is required for targeting to sites of actin polarization,” *J Cell Biol*, vol. 141, no. 6, pp. 1357–70, 1998.

[123] W. Xu, S. C. Harrison, and M. J. Eck, “Three-dimensional structure of the tyrosine kinase c-Src,” *Nature*, vol. 385, no. 6617, pp. 595–602, 1997.

[124] F. Sicheri, I. Moarefi, and J. Kuriyan, “Crystal structure of the Src family tyrosine kinase Hck,” *Nature*, vol. 385, no. 6617, pp. 602–9, 1997.

[125] Y. Wu, S. D. Spencer, and L. A. Lasky, “Tyrosine phosphorylation regulates the SH3-mediated binding of the Wiskott-Aldrich syndrome protein to PSTPIP, a cytoskeletal-associated protein,” *Journal of Biological Chemistry*, vol. 273, no. 10, pp. 5765–70, 1998.

[126] H. Zhao, S. Okada, J. E. Pessin, and G. A. Koretzky, “Insulin receptor-mediated dissociation of Grb2 from Sos involves phosphorylation of Sos by kinase(s) other than extracellular signal-regulated kinase,” *Journal of Biological Chemistry*, vol. 273, no. 20, pp. 12061–7, 1998.

[127] A. R. Comer, S. M. Ahern-Djamali, J. L. Juang, P. D. Jackson, and F. M. Hoffmann, “Phosphorylation of Enabled by the Drosophila Abelson tyrosine kinase regulates the in vivo function and protein-protein interactions of Enabled,” *Mol Cell Biol*, vol. 18, no. 1, pp. 152–60, 1998.

[128] M. C. Parrini and B. J. Mayer, “Engineering temperature-sensitive SH3 domains,” *Chem Biol*, vol. 6, no. 10, pp. 679–87, 1999.

[129] J. C. Arevalo, D. B. Pereira, H. Yano, K. K. Teng, and M. V. Chao, “Identification of a switch in neurotrophin signaling by selective tyrosine phosphorylation,” *Journal of Biological Chemistry*, vol. 281, no. 2, pp. 1001–7, 2006.

[130] C. Wu, M. H. Ma, K. R. Brown, M. Geisler, L. Li, E. Tzeng, C. Y. H. Jia, I. Jurisica, and S. S.-C. Li, “Systematic identification of SH3 domain-mediated human protein-protein interactions by peptide array target screening,” *Proteomics*, vol. 7, no. 11, pp. 1775–85, 2007.

[131] A. Sanjay, T. Miyazaki, C. Itzstein, E. Purev, W. C. Horne, and R. Baron, “Identification and functional characterization of an Src homology domain 3 domain-binding site on Cbl,” *FEBS J*, vol. 273, no. 23, pp. 5442–56, 2006.

[132] S. D. Dimitratos, D. F. Woods, D. G. Stathakis, and P. J. Bryant, “Signaling pathways are focused at specialized regions of the plasma membrane by scaffolding proteins of the MAGUK family,” *Bioessays*, vol. 21, no. 11, pp. 912–21, 1999.

[133] H. Shin, Y. P. Hsueh, F. C. Yang, E. Kim, and M. Sheng, “An intramolecular interaction between Src homology 3 domain and guanylate kinase-like domain required for channel clustering by postsynaptic density-95/SAP90,” *J Neurosci*, vol. 20, no. 10, pp. 3580–7, 2000.

[134] R. Li, “Bee1, a yeast protein with homology to Wiscott-Aldrich syndrome protein, is critical for the assembly of cortical actin cytoskeleton,” *J Cell Biol*, vol. 136, no. 3, pp. 649–58, 1997.

[135] S. N. Naqvi, R. Zahn, D. A. Mitchell, B. J. Stevenson, and A. L. Munn, “The WASp homologue Las17p functions with the WIP homologue End5p/verprolin and is essential for endocytosis in yeast,” *Curr Biol*, vol. 8, no. 17, pp. 959–62, 1998.

[136] M. Evangelista, B. M. Klebl, A. H. Tong, B. A. Webb, T. Leeuw, E. Leberer, M. Whiteway, D. Y. Thomas, and C. Boone, “A role for myosin-I in actin assembly through interactions with Vrp1p, Bee1p, and the Arp2/3 complex,” *J Cell Biol*, vol. 148, no. 2, pp. 353–62, 2000.

[137] H. V. Goodson, B. L. Anderson, H. M. Warrick, L. A. Pon, and J. A. Spudich, “Synthetic lethality screen identifies a novel yeast myosin I gene (MYO5): myosin I proteins are required for polarization of the actin cytoskeleton,” *J Cell Biol*, vol. 133, no. 6, pp. 1277–91, 1996.

[138] M. I. Geli, R. Lombardi, B. Schmelzl, and H. Riezman, “An intact SH3 domain is required for myosin I-induced actin polymerization,” *EMBO J*, vol. 19, no. 16, pp. 4281–91, 2000.

[139] H.-J. Lee and J. J. Zheng, “PDZ domains and their binding partners: structure, specificity, and modification,” *Cell Commun Signal*, vol. 8, p. 8, 2010.

[140] H. J. Chung, Y. H. Huang, L.-F. Lau, and R. L. Huganir, “Regulation of the NMDA receptor complex and trafficking by activity-dependent phosphorylation of the NR2B subunit PDZ ligand,” *J Neurosci*, vol. 24, no. 45, pp. 10248–59, 2004.

[141] P. Steiner, M. J. Higley, W. Xu, B. L. Czervionke, R. C. Malenka, and B. L. Sabatini, “Destabilization of the postsynaptic density by PSD-95 serine 73 phosphorylation inhibits spine growth and synaptic plasticity,” *Neuron*, vol. 60, no. 5, pp. 788–802, 2008.

[142] E. Akiva, G. Friedlander, Z. Itzhaki, and H. Margalit, “A dynamic view of domain-motif interactions,” *PLoS Comput Biol*, vol. 8, no. 1, p. e1002341, 2012.

[143] L. C. J. van den Berk, E. Landi, T. Walma, G. W. Vuister, L. Dente, and W. J. A. J. Hendriks, “An allosteric intramolecular PDZ-PDZ interaction modulates PTP-BL PDZ2 binding specificity,” *Biochemistry*, vol. 46, no. 47, pp. 13629–37, 2007.

[144] Y. Chen, R. Sheng, M. Kallberg, A. Silkov, M. P. Tun, N. Bhardwaj, S. Kurilova, R. A. Hall, B. Honig, H. Lu, and W. Cho, “Genome-wide functional annotation of dual-specificity protein- and lipid-binding modules that regulate protein interactions,” *Mol Cell*, vol. 46, no. 2, pp. 226–37, 2012.

[145] S. Bhattacharya, Z. Dai, J. Li, S. Baxter, D. J. E. Callaway, D. Cowburn, and Z. Bu, “A conformational switch in the scaffolding protein NHERF1 controls autoinhibition and complex formation,” *Journal of Biological Chemistry*, vol. 285, no. 13, pp. 9981–94, 2010.

[146] J.-F. Long, W. Feng, R. Wang, L.-N. Chan, F. C. F. Ip, J. Xia, N. Y. Ip, and M. Zhang, "Autoinhibition of X11/Mint scaffold proteins revealed by the closed conformation of the PDZ tandem," *Nat Struct Mol Biol*, vol. 12, no. 8, pp. 722–8, 2005.

[147] S. M. Kaech, C. W. Whitfield, and S. K. Kim, "The LIN-2/LIN-7/LIN-10 complex mediates basolateral membrane localization of the C. elegans EGF receptor LET-23 in vulval epithelial cells," *Cell*, vol. 94, no. 6, pp. 761–71, 1998.

[148] S. Maudsley, A. M. Zamah, N. Rahman, J. T. Blitzer, L. M. Luttrell, R. J. Lefkowitz, and R. A. Hall, "Platelet-derived growth factor receptor association with Na(+)/H(+) exchanger regulatory factor potentiates receptor activity," *Mol Cell Biol*, vol. 20, no. 22, pp. 8352–63, 2000.

[149] D. Bilder, M. Li, and N. Perrimon, "Cooperative regulation of cell polarity and growth by Drosophila tumor suppressors," *Science*, vol. 289, no. 5476, pp. 113–6, 2000.

[150] J. D. Hildebrand and P. Soriano, "Shroom, a PDZ domain-containing actin-binding protein, is required for neural tube morphogenesis in mice," *Cell*, vol. 99, no. 5, pp. 485–97, 1999.

[151] G. Caruana and A. Bernstein, "Craniofacial dysmorphogenesis including cleft palate in mice with an insertional mutation in the discs large gene," *Mol Cell Biol*, vol. 21, no. 5, pp. 1475–83, 2001.

[152] J. N. Murdoch, D. J. Henderson, K. Doudney, C. Gaston-Massuet, H. M. Phillips, C. Paternotte, R. Arkell, P. Stanier, and A. J. Copp, "Disruption of scribble (Scrb1) causes severe neural tube defects in the circletail mouse," *Hum Mol Genet*, vol. 12, no. 2, pp. 87–98, 2003.

[153] M.-J. Santoni, P. Pontarotti, D. Birnbaum, and J.-P. Borg, "The LAP family: a phylogenetic point of view," *Trends in Genetics*, vol. 18, no. 10, pp. 494–7, 2002.

[154] B. A. Liu, B. W. Engelmann, and P. D. Nash, "High-throughput analysis of peptide-binding modules," *Proteomics*, vol. 12, no. 10, pp. 1527–46, 2012.

[155] R. Volkmer, V. Tapia, and C. Landgraf, "Synthetic peptide arrays for investigating protein interaction domains," *FEBS Lett*, vol. 586, no. 17, pp. 2780–6, 2012.

[156] M. Rodriguez, S. S.-C. Li, J. W. Harper, and Z. Songyang, "An oriented peptide array library (OPAL) strategy to study protein-protein interactions," *Journal of Biological Chemistry*, vol. 279, no. 10, pp. 8802–7, 2004.

[157] K. Machida, C. M. Thompson, K. Dierck, K. Jablonowski, S. Karkkainen, B. Liu, H. Zhang, P. D. Nash, D. K. Newman, P. Nollau, T. Pawson, G. H. Renkema, K. Saksela, M. R. Schiller, D.-G. Shin, and B. J. Mayer, "High-throughput phosphotyrosine profiling using SH2 domains," *Mol Cell*, vol. 26, no. 6, pp. 899–915, 2007.

[158] R. Frank, W. Heikens, G. Heisterberg-Moutsis, and H. Blocker, "A new general approach for the simultaneous chemical synthesis of large numbers of oligonucleotides: segmental solid supports," *Nucleic Acids Res*, vol. 11, no. 13, pp. 4365–77, 1983.

[159] R. Frank, "Spot-synthesis: an easy technique for the positionally addressable, parallel chemical synthesis on a membrane support," *Tetrahedron*, vol. 48, pp. 9217–9232, 1992.

[160] S. P. Fodor, J. L. Read, M. C. Pirrung, L. Stryer, A. T. Lu, and D. Solas, "Light-directed, spatially addressable parallel chemical synthesis," *Science*, vol. 251, no. 4995, pp. 767–73, 1991.

[161] R. Frank, "The SPOT-synthesis technique. Synthetic peptide arrays on membrane supports–principles and applications," *J Immunol Methods*, vol. 267, no. 1, pp. 13–26, 2002.

[162] M. Tinti, L. Kiemer, S. Costa, M. L. Miller, F. Sacco, J. V. Olsen, M. Carducci, S. Paoluzi, F. Langone, C. T. Workman, N. Blom, K. Machida, C. M. Thompson, M. Schutkowski, S. Brunak, M. Mann, B. J. Mayer, L. Castagnoli, and G. Cesareni, "The SH2 domain interaction landscape," *Cell Rep*, vol. 3, no. 4, pp. 1293–305, 2013.

[163] P. Boisguerin, R. Leben, B. Ay, G. Radziwill, K. Moelling, L. Dong, and R. Volkmer-Engert, "An improved method for the synthesis of cellulose membrane-bound peptides with free C termini is useful for PDZ domain binding studies," *Chem Biol*, vol. 11, no. 4, pp. 449–59, 2004.

[164] M. A. Stiffler, J. R. Chen, V. P. Grantcharova, Y. Lei, D. Fuchs, J. E. Allen, L. A. Zaslavskaia, and G. MacBeath, "PDZ domain binding selectivity is optimized across the mouse proteome," *Science*, vol. 317, no. 5836, pp. 364–9, 2007.

[165] R. B. Jones, A. Gordus, J. A. Krall, and G. MacBeath, "A quantitative protein interaction network for the ErbB receptors using protein microarrays," *Nature*, vol. 439, no. 7073, pp. 168–74, 2006.

[166] B. W. Engelmann, Y. Kim, M. Wang, B. Peters, R. S. Rock, and P. D. Nash, "The development and application of a quantitative peptide microarray based approach to protein interaction domain specificity space," *Mol Cell Proteomics*, 2014.

[167] S. Hu, Z. Xie, J. Qian, S. Blackshaw, and H. Zhu, "Functional protein microarray technology," *Wiley Interdiscip Rev Syst Biol Med*, vol. 3, no. 3, pp. 255–68, 2011.

[168] G. P. Smith, "Filamentous fusion phage: novel expression vectors that display cloned antigens on the virion surface," *Science*, vol. 228, no. 4705, pp. 1315–7, 1985.

[169] G. P. Smith and V. A. Petrenko, "Phage Display," *Chem Rev*, vol. 97, no. 2, pp. 391–410, 1997.

[170] S. S. Sidhu, W. J. Fairbrother, and K. Deshayes, "Exploring protein-protein interactions with phage display," *Chembiochem*, vol. 4, no. 1, pp. 14–25, 2003.

[171] G. Fuh, M. T. Pisabarro, Y. Li, C. Quan, L. A. Lasky, and S. S. Sidhu, "Analysis of PDZ domain-ligand interactions using carboxyl-terminal phage display," *Journal of Biological Chemistry*, vol. 275, no. 28, pp. 21486–91, 2000.

[172] W. Mandecki, Y. C. Chen, and N. Grihalde, "A mathematical model for biopanning (affinity selection) using peptide libraries on filamentous phage," *J Theor Biol*, vol. 176, no. 4, pp. 523–30, 1995.

[173] R. Tonikian, X. Xin, C. P. Toret, D. Gfeller, C. Landgraf, S. Panni, S. Paoluzi, L. Castagnoli, B. Currell, S. Seshagiri, H. Yu, B. Winsor, M. Vidal, M. B. Gerstein, G. D. Bader, R. Volkmer,

G. Cesareni, D. G. Drubin, P. M. Kim, S. S. Sidhu, and C. Boone, "Bayesian modeling of the yeast SH3 domain interactome predicts spatiotemporal dynamics of endocytosis proteins," *PLoS Biol*, vol. 7, no. 10, p. e1000218, 2009.

[174] J. Teyra, S. S. Sidhu, and P. M. Kim, "Elucidation of the binding preferences of peptide recognition modules: SH3 and PDZ domains," *FEBS Lett*, 2012.

[175] T. Mi, J. C. Merlin, S. Deverasetty, M. R. Gryk, T. J. Bill, A. W. Brooks, L. Y. Lee, V. Rathnayake, C. A. Ross, D. P. Sargeant, C. L. Strong, P. Watts, S. Rajasekaran, and M. R. Schiller, "Minimotif Miner 3.0: database expansion and significantly improved reduction of false-positive predictions from consensus sequences," *Nucleic Acids Res*, vol. 40, no. Database issue, pp. D252–60, 2012.

[176] H. Dinkel, K. Van Roey, S. Michael, N. E. Davey, R. J. Weatheritt, D. Born, T. Speck, D. Kruger, G. Grebnev, M. Kuban, M. Strumillo, B. Uyar, A. Budd, B. Altenberg, M. Seiler, L. B. Chemes, J. Glavina, I. E. Sanchez, F. Diella, and T. J. Gibson, "The eukaryotic linear motif resource ELM: 10 years and counting," *Nucleic Acids Res*, vol. 42, no. Database issue, pp. D259–66, 2014.

[177] L. Gold, D. Pribnow, T. Schneider, S. Shinedling, B. S. Singer, and G. Stormo, "Translational initiation in prokaryotes," *Annual review of microbiology*, vol. 35, pp. 365–403, 1981.

[178] G. D. Stormo, T. D. Schneider, and L. M. Gold, "Characterization of translational initiation sites in E. coli," *Nucleic Acids Res*, vol. 10, no. 9, pp. 2971–96, 1982.

[179] T. D. Schneider and R. M. Stephens, "Sequence logos: a new way to display consensus sequences," *Nucleic Acids Res*, vol. 18, no. 20, pp. 6097–100, 1990.

[180] G. E. Crooks, G. Hon, J.-M. Chandonia, and S. E. Brenner, "WebLogo: a sequence logo generator," *Genome Res*, vol. 14, no. 6, pp. 1188–90, 2004.

[181] J. C. Obenauer, L. C. Cantley, and M. B. Yaffe, "Scansite 2.0: Proteome-wide prediction of cell signaling interactions using short sequence motifs," *Nucleic Acids Res*, vol. 31, no. 13, pp. 3635–41, 2003.

[182] M. B. Yaffe, G. G. Leparc, J. Lai, T. Obata, S. Volinia, and L. C. Cantley, "A motif-based profile scanning approach for genome-wide prediction of signaling pathways," *Nat Biotechnol*, vol. 19, no. 4, pp. 348–53, 2001.

[183] L. Li, C. Wu, H. Huang, K. Zhang, J. Gan, and S. S.-C. Li, "Prediction of phosphotyrosine signaling networks using a scoring matrix-assisted ligand identification approach," *Nucleic Acids Res*, vol. 36, no. 10, pp. 3263–73, 2008.

[184] E. Zaslavsky, P. Bradley, and C. Yanover, "Inferring PDZ domain multi-mutant binding preferences from single-mutant data," *PLoS One*, vol. 5, no. 9, p. e12787, 2010.

[185] H. Y. K. Lam, P. M. Kim, J. Mok, R. Tonikian, S. S. Sidhu, B. E. Turk, M. Snyder, and M. B. Gerstein, "MOTIPS: automated motif analysis for predicting targets of modular protein domains," *BMC Bioinformatics*, vol. 11, p. 243, 2010.

[186] B. Brannetti and M. Helmer-Citterich, “iSPOT: A web tool to infer the interaction specificity of families of protein modules,” *Nucleic Acids Res*, vol. 31, no. 13, pp. 3709–11, 2003.

[187] D. Gfeller, “Uncovering new aspects of protein interactions through analysis of specificity landscapes in peptide recognition domains,” *FEBS Lett*, vol. 586, no. 17, pp. 2764–72, 2012.

[188] D. Gfeller, F. Butty, M. Wierzbicka, E. Verschueren, P. Vanhee, H. Huang, A. Ernst, N. Dar, I. Stagljar, L. Serrano, S. S. Sidhu, G. D. Bader, and P. M. Kim, “The multiple-specificity landscape of modular peptide recognition domains,” *Mol Syst Biol*, vol. 7, p. 484, 2011.

[189] L. Kaustov, H. Ouyang, M. Amaya, A. Lemak, N. Nady, S. Duan, G. A. Wasney, Z. Li, M. Vedadi, M. Schapira, J. Min, and C. H. Arrowsmith, “Recognition and specificity determinants of the human cbx chromodomains,” *Journal of Biological Chemistry*, vol. 286, no. 1, pp. 521–9, 2011.

[190] S. Panni, L. Montecchi-Palazzi, L. Kiemer, A. Cabibbo, S. Paoluzi, E. Santonico, C. Landgraf, R. Volkmer-Engert, A. Bachi, L. Castagnoli, and G. Cesareni, “Combining peptide recognition specificity and context information for the prediction of the 14-3-3-mediated interactome in S. cerevisiae and H. sapiens,” *Proteomics*, vol. 11, no. 1, pp. 128–43, 2011.

[191] M. L. Miller, L. J. Jensen, F. Diella, C. Jorgensen, M. Tinti, L. Li, M. Hsiung, S. A. Parker, J. Bordeaux, T. Sicheritz-Ponten, M. Olhovsky, A. Pasculescu, J. Alexander, S. Knapp, N. Blom, P. Bork, S. Li, G. Cesareni, T. Pawson, B. E. Turk, M. B. Yaffe, S. Brunak, and R. Linding, “Linear motif atlas for phosphorylation-dependent signaling,” *Sci Signal*, vol. 1, no. 35, p. ra2, 2008.

[192] T. Kim, M. S. Tyndel, H. Huang, S. S. Sidhu, G. D. Bader, D. Gfeller, and P. M. Kim, “MUSI: an integrated system for identifying multiple specificity from very large peptide or nucleic acid data sets,” *Nucleic Acids Res*, 2011.

[193] T. L. Bailey and C. Elkan, “The value of prior knowledge in discovering motifs with MEME,” in *Proc. of the 3th Int. Conf. on Intelligent Systems for Molecular Biology (ISMB’95)*, vol. 3, pp. 21–9, 1995.

[194] K. Luck and G. Trave, “Phage display can select over-hydrophobic sequences that may impair prediction of natural domain-peptide interactions,” *Bioinformatics*, vol. 27, no. 7, pp. 899–902, 2011.

[195] K. Katoh, K. Misawa, K.-i. Kuma, and T. Miyata, “MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform,” *Nucleic Acids Res*, vol. 30, no. 14, pp. 3059–66, 2002.

[196] L. Licata, L. Briganti, D. Peluso, L. Perfetto, M. Iannuccelli, E. Galeota, F. Sacco, A. Palma, A. P. Nardozza, E. Santonico, L. Castagnoli, and G. Cesareni, “MINT, the molecular interaction database: 2012 update,” *Nucleic Acids Res*, vol. 40, no. Database issue, pp. D857–61, 2012.

[197] A. Y. Ng and M. I. Jordan, “On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes,” in *NIPS*, pp. 841–848, 2001.

[198] X. Shao, C. S. H. Tan, C. Voss, S. S. C. Li, N. Deng, and G. D. Bader, "A regression framework incorporating quantitative and negative interaction data improves quantitative prediction of PDZ domain-peptide interaction from primary sequence," *Bioinformatics*, vol. 27, no. 3, pp. 383–90, 2011.

[199] S. Hui and G. D. Bader, "Proteome scanning to predict PDZ domain interactions using support vector machines," *BMC Bioinformatics*, vol. 11, p. 507, 2010.

[200] L. Li, B. Zhao, J. Du, K. Zhang, C. X. Ling, and S. S.-C. Li, "DomPep–a general method for predicting modular domain-mediated protein-protein interactions," *PLoS One*, vol. 6, no. 10, p. e25528, 2011.

[201] H.-S. Eo, S. Kim, H. Koo, and W. Kim, "A machine learning based method for the prediction of G protein-coupled receptor-binding PDZ domain proteins," *Mol Cells*, vol. 27, no. 6, pp. 629–34, 2009.

[202] S. Kalyoncu, O. Keskin, and A. Gursoy, "Interaction prediction and classification of PDZ domains," *BMC Bioinformatics*, vol. 11, p. 357, 2010.

[203] S. Hui, X. Xing, and G. D. Bader, "Predicting PDZ domain mediated protein interactions from structure," *BMC Bioinformatics*, vol. 14, p. 27, 2013.

[204] J. C. Hawkins, H. Zhu, J. Teyra, and M. T. Pisabarro, "Reduced False Positives in PDZ Binding Prediction using Sequence and Structural Descriptors," *IEEE/ACM Trans Comput Biol Bioinform*, 2012.

[205] E. Ferraro, A. Via, G. Ausiello, and M. Helmer-Citterich, "A novel structure-based encoding for machine-learning applied to the inference of SH3 domain specificity," *Bioinformatics*, vol. 22, no. 19, pp. 2333–9, 2006.

[206] E. Ferraro, D. Peluso, A. Via, G. Ausiello, and M. Helmer-Citterich, "SH3-Hunter: discovery of SH3 domain interaction sites in proteins," *Nucleic Acids Res*, vol. 35, no. Web Server issue, pp. W451–4, 2007.

[207] F. E. Witten IH, *Data Mining: Practical machine learning tools and techniques.* Morgan Kaufmann, San Francisco, 2005.

[208] D. J. Reiss and B. Schwikowski, "Predicting protein-peptide interactions via a network-based motif sampler," *Bioinformatics*, vol. 20 Suppl 1, pp. I274–I282, 2004.

[209] L. Zhang, C. Shao, D. Zheng, and Y. Gao, "An integrated machine learning system to computationally screen protein databases for protein binding peptide ligands," *Mol Cell Proteomics*, vol. 5, no. 7, pp. 1224–32, 2006.

[210] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, pp. 1263–1284, 2009.

[211] F. Provost, "Machine learning from imbalanced data sets 101," in *Proceedings of the AAAI-2000 Workshop on Imbalanced Data Sets*, 2000.

[212] Jo and Japkowicz, “Class imbalances versus small disjuncts,” in *ACM SIGKDD Explorations Newsletter*, 2004.

[213] N. Chawla, K. Bowyer, L. Hall, and W. P. Kegelmeyer, “Smote: synthetic minority oversampling technique,” *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, Jan 2002.

[214] X. Zhu, “Semi-supervised learning literature survey,” Tech. Rep. 1530, Computer Sciences, University of Wisconsin-Madison, 2005.

[215] F. Cozman, I. Cohen, and M. Cirelo, “Semi-supervised learning of mixture models and bayesian networks,” in *Proceedings of the Twentieth International Conference of Machine Learning*, pp. 1–8, 2003.

[216] A. Ben-Hur and W. S. Noble, “Choosing negative examples for the prediction of protein-protein interactions,” *BMC Bioinformatics*, vol. 7 Suppl 1, p. S2, 2006.

[217] S. L. Lo, C. Z. Cai, Y. Z. Chen, and M. C. M. Chung, “Effect of training datasets on support vector machine prediction of protein-protein interactions,” *Proteomics*, vol. 5, no. 4, pp. 876–84, 2005.

[218] J. K. Lee, T. Moon, M. W. Chi, J.-S. Song, Y.-S. Choi, and C. N. Yoon, “An investigation of phosphopeptide binding to sh2 domain.,” *Biochem Biophys Res Commun*, vol. 306, pp. 225–230, Jun 2003.

[219] I. E. Sanchez, P. Beltrao, F. Stricher, J. Schymkowitz, J. Ferkinghoff-Borg, F. Rousseau, and L. Serrano, “Genome-wide prediction of SH2 domain targets using structural information and the FoldX algorithm,” *PLoS Comput Biol*, vol. 4, no. 4, p. e1000052, 2008.

[220] C. A. Smith and T. Kortemme, “Structure-based prediction of the peptide sequence space recognized by natural and synthetic PDZ domains,” *J Mol Biol*, vol. 402, no. 2, pp. 460–74, 2010.

[221] K. Kaufmann, N. Shen, L. Mizoue, and J. Meiler, “A physical model for PDZ-domain/peptide interactions,” *J Mol Model*, vol. 17, no. 2, pp. 315–24, 2011.

[222] T. Hou, N. Li, Y. Li, and W. Wang, “Characterization of domain-peptide interaction interface: prediction of SH3 domain-mediated protein-protein interaction network in yeast by generic structure-based models,” *J Proteome Res*, vol. 11, no. 5, pp. 2982–95, 2012.

[223] G. Fernandez-Ballester, P. Beltrao, J. M. Gonzalez, Y.-H. Song, M. Wilmanns, A. Valencia, and L. Serrano, “Structure-based prediction of the Saccharomyces cerevisiae SH3-ligand interactions,” *J Mol Biol*, vol. 388, no. 4, pp. 902–16, 2009.

[224] D. A. Henriques, J. E. Ladbury, and R. M. Jackson, “Comparison of binding energies of SrcSH2-phosphotyrosyl peptides with structure-based prediction using surface area based empirical parameterization,” *Protein Sci*, vol. 9, no. 10, pp. 1975–85, 2000.

[225] W. A. McLaughlin, T. Hou, and W. Wang, “Prediction of binding sites of peptide recognition domains: an application on Grb2 and SAP SH2 domains,” *J Mol Biol*, vol. 357, no. 4, pp. 1322–34, 2006.

[226] A. Suenaga, M. Hatakeyama, M. Ichikawa, X. Yu, N. Futatsugi, T. Narumi, K. Fukui, T. Terada, M. Taiji, M. Shirouzu, S. Yokoyama, and A. Konagaya, "Molecular dynamics, free energy, and SPR analyses of the interactions between the SH2 domain of Grb2 and ErbB phosphotyrosyl peptides," *Biochemistry*, vol. 42, no. 18, pp. 5195–200, 2003.

[227] T. Hou, K. Chen, W. A. McLaughlin, B. Lu, and W. Wang, "Computational analysis and prediction of the binding motif and protein interacting partners of the Abl SH3 domain," *PLoS Comput Biol*, vol. 2, no. 1, p. e1, 2006.

[228] C. A. King and P. Bradley, "Structure-based prediction of protein-peptide specificity in Rosetta," *Proteins*, vol. 78, no. 16, pp. 3437–49, 2010.

[229] A. M. Wollacott and J. R. Desjarlais, "Virtual interaction profiles of proteins," *J Mol Biol*, vol. 313, no. 2, pp. 317–42, 2001.

[230] Z. Wunderlich and L. A. Mirny, "Using genome-wide measurements for computational prediction of SH2-peptide interactions," *Nucleic Acids Res*, vol. 37, no. 14, pp. 4629–41, 2009.

[231] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, pp. 273–297, 1995.

[232] T. Joachims, *Making large-scale SVM learning practical, in Advanced in Kernel Methods-Support Vector Learning (ikopf, B., Burges, C., Smola, A., eds) pp. 169-184.* MIT Press, Cambridge, MA, 1999.

[233] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Springer, second ed., 2008.

[234] N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods.* Cambridge University Press, 2010.

[235] P. Baldi, S. Brunak, Y. Chauvin, C. A. Andersen, and H. Nielsen, "Assessing the accuracy of prediction algorithms for classification: an overview," *Bioinformatics*, vol. 16, no. 5, pp. 412–24, 2000.

[236] F. Costa and K. D. Grave, "Fast neighborhood subgraph pairwise distance kernel," in *Proceedings of the 26 th International Conference on Machine Learning*, pp. 255–262, Omnipress, 2010.

[237] J. Kyte and R. F. Doolittle, "A simple method for displaying the hydropathic character of a protein," *J Mol Biol*, vol. 157, no. 1, pp. 105–32, 1982.

[238] B. E. Boser, I. M. Guyon, and V. N. Vapnik, "A training algorithm for optimal margin classifiers," in *Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory*, pp. 144–152, ACM Press, 1992.

[239] M. A. Aizerman, E. M. Braverman, and L. I. Rozonoer, "Theoretical foundations of the potential function method in pattern recognition learning," in *Automat. Remote Contr.*, vol. 25, pp. 917 – 936, 1964.

[240] D. Haussler, “Convolution kernels on discrete structures,” Technical Report UCS-CRL-99-10, University of California at Santa Cruz, Santa Cruz, CA, USA, 1999.

[241] E. M. Luks, “Isomorphism of graphs of bounded valence can be tested in polynomial time,” *J. Comput. Syst. Sci*, vol. 25, pp. 42–65, 1982.

[242] B. D. McKay, “Practical graph isomorphism,” *Congressus Numerantium*, vol. 30, pp. 45–87, 1981.

[243] X. Yan and J. Han., “gSpan: Graph-based substructure pattern mining,” in *Proc. 2002 Int. Conf. Data Mining (ICDM ´02)*, pp. 721–724, 2002.

[244] S. Sorlin and C. Solnon, “A parametric filtering algorithm for the graph isomorphism problem,” *Constraints*, vol. 13, pp. 518–537, 2008.

[245] I. Damgård, “A design principle for hash functions,” in *Advances in Cryptology-CRYPTO ´89 Proceedings*, pp. 416–427, Springer, 1990.

[246] S. Heyne, F. Costa, D. Rose, and R. Backofen, “GraphClust: alignment-free structural clustering of local RNA secondary structures,” *Bioinformatics*, vol. 28, no. 12, pp. i224–i232, 2012.

[247] P. Rice, I. Longden, and A. Bleasby, “EMBOSS: the European Molecular Biology Open Software Suite,” *Trends in Genetics*, vol. 16, no. 6, pp. 276–7, 2000.

[248] T. Bailey and C. Elkan, “The value of prior knowledge in discovering motifs with meme,” *Proc Int Conf Intell Syst Mol Biol*, vol. 3, pp. 21–9, 1995.

[249] T. L. Bailey and M. Gribskov, “Combining evidence using p-values: application to sequence homology searches,” *Bioinformatics*, vol. 14, no. 1, pp. 48–54, 1998.

[250] M. Kasowski, F. Grubert, C. Heffelfinger, M. Hariharan, A. Asabere, S. M. Waszak, L. Habegger, J. Rozowsky, M. Shi, A. E. Urban, M.-Y. Hong, K. J. Karczewski, W. Huber, S. M. Weissman, M. B. Gerstein, J. O. Korbel, and M. Snyder, “Variation in transcription factor binding among humans,” *Science*, vol. 328, no. 5975, pp. 232–5, 2010.

[251] B. Schölkopf, J. C. Platt, J. C. Shawe-Taylor, A. J. Smola, and R. C. Williamson, “Estimating the support of a high-dimensional distribution,” *Neural Comput.*, vol. 13, pp. 1443–1471, July 2001.

[252] M. Culp and G. Michailidis, “An iterative algorithm for extending learners to a semisupervised setting,” in *The 2007 Joint Statistical Meetings (JSM*, 2007.

[253] R. Caruana, “Multitask learning,” *Machine Learning*, vol. 28, pp. 41–75, 1997.

[254] S. van Dongen, *Graph Clustering by Flow Simulation.* PhD thesis, University of Utrecht, The Netherlands, 2000.

[255] A. J. Enright, S. Van Dongen, and C. A. Ouzounis, “An efficient algorithm for large-scale detection of protein families,” *Nucleic Acids Res*, vol. 30, no. 7, pp. 1575–84, 2002.

[256] S. B. Needleman and C. D. Wunsch, “A general method applicable to the search for similarities in the amino acid sequence of two proteins,” *J Mol Biol*, vol. 48, no. 3, pp. 443–53, 1970.

[257] J. Shawe-Taylor and N. Cristianini, *Kernel methods for pattern analysis*. Cambridge university press, 2004.

[258] M. Andreatta, O. Lund, and M. Nielsen, “Simultaneous alignment and clustering of peptide data using a Gibbs sampling approach,” *Bioinformatics*, vol. 29, no. 1, pp. 8–14, 2013.

[259] A. V. Persikov, R. Osada, and M. Singh, “Predicting DNA recognition by Cys2His2 zinc finger proteins,” *Bioinformatics*, vol. 25, no. 1, pp. 22–9, 2009.

[260] F. Diella, C. M. Gould, C. Chica, A. Via, and T. J. Gibson, “Phospho.ELM: a database of phosphorylation sites–update 2008,” *Nucleic Acids Res*, vol. 36, no. Database issue, pp. D240–4, 2008.

[261] R. Tonikian, Y. Zhang, C. Boone, and S. S. Sidhu, “Identifying specificity profiles for peptide recognition modules from phage-displayed peptide libraries,” *Nat Protoc*, vol. 2, no. 6, pp. 1368–86, 2007.

[262] C. Landgraf, S. Panni, L. Montecchi-Palazzi, L. Castagnoli, J. Schneider-Mergener, R. Volkmer-Engert, and G. Cesareni, “Protein interaction networks by proteome peptide scanning,” *PLoS Biol*, vol. 2, no. 1, p. E14, 2004.

[263] L. Bottou and O. Bousquet, “The tradeoffs of large scale learning,” in *Advances in Neural Information Processing Systems* (J. Platt, D. Koller, Y. Singer, and S. Roweis, eds.), vol. 20, pp. 161–168, NIPS Foundation (http://books.nips.cc), 2008.

[264] T. Beuming, L. Skrabanek, M. Y. Niv, P. Mukherjee, and H. Weinstein, “PDZBase: a protein-protein interaction database for PDZ-domains,” *Bioinformatics*, vol. 21, no. 6, pp. 827–8, 2005.

[265] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, “The protein data bank,” *Nucleic Acids Res*, vol. 28, no. 1, pp. 235–42, 2000.

[266] P. V. Hornbeck, J. M. Kornhauser, S. Tkachev, B. Zhang, E. Skrzypek, B. Murray, V. Latham, and M. Sullivan, “PhosphoSitePlus: a comprehensive resource for investigating the structure and function of experimentally determined post-translational modifications in man and mouse,” *Nucleic Acids Res*, vol. 40, no. Database issue, pp. D261–70, 2012.

[267] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock, “Gene ontology: tool for the unification of biology. The Gene Ontology Consortium,” *Nat Genet*, vol. 25, no. 1, pp. 25–9, 2000.

[268] C.-A. Tanase, “Histidine domain-protein tyrosine phosphatase interacts with Grb2 and GrpL,” *PLoS One*, vol. 5, no. 12, p. e14339, 2010.

[269] Z. Dosztanyi, V. Csizmok, P. Tompa, and I. Simon, “IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content,” *Bioinformatics*, vol. 21, no. 16, pp. 3433–4, 2005.

[270] H. Faderl, S. Kantarjian and M. Talpaz, “Chronic myelogenous leukemia: Update on biology and treatment,” *Oncology (Williston Park)*, vol. 13, pp. 169–180, 1999.

[271] F. U. Wohrle, S. Halbach, K. Aumann, S. Schwemmers, S. Braun, P. Auberger, D. Schramek, J. M. Penninger, S. Lassmann, M. Werner, C. F. Waller, H. L. Pahl, R. Zeiser, R. J. Daly, and T. Brummer, “Gab2 signaling in chronic myeloid leukemia cells confers resistance to multiple Bcr-Abl inhibitors,” *Leukemia*, 2012.

[272] C. Preisinger, J. P. Schwarz, O. B. Bleijerveld, E. Corradini, P. J. Muller, K. I. Anderson, W. Kolch, A. Scholten, and A. J. R. Heck, “Imatinib-dependent tyrosine phosphorylation profiling of Bcr-Abl-positive chronic myeloid leukemia cells,” *Leukemia*, vol. 27, no. 3, pp. 743–6, 2013.

[273] A. Hamilton, L. Elrick, S. Myssina, M. Copland, H. Jorgensen, J. V. Melo, and T. Holyoake, “BCR-ABL activity and its response to drugs can be determined in CD34+ CML stem cells by CrkL phosphorylation status using flow cytometry,” *Leukemia*, vol. 20, no. 6, pp. 1035–9, 2006.

[274] C. I. Smith, T. C. Islam, P. T. Mattsson, A. J. Mohamed, B. F. Nore, and M. Vihinen, “The Tec family of cytoplasmic tyrosine kinases: mammalian Btk, Bmx, Itk, Tec, Txk and homologs in other species,” *Bioessays*, vol. 23, no. 5, pp. 436–46, 2001.

[275] R. Marone, V. Cmiljanovic, B. Giese, and M. P. Wymann, “Targeting phosphoinositide 3-kinase: moving towards therapy,” *Biochim Biophys Acta*, vol. 1784, no. 1, pp. 159–85, 2008.

[276] M. G. Tomlinson, V. L. Heath, C. W. Turck, S. P. Watson, and A. Weiss, “SHIP family inositol phosphatases interact with and negatively regulate the Tec tyrosine kinase,” *Journal of Biological Chemistry*, vol. 279, no. 53, pp. 55089–96, 2004.

[277] M. Huber, C. D. Helgason, J. E. Damen, M. Scheid, V. Duronio, L. Liu, M. D. Ware, R. K. Humphries, and G. Krystal, “The role of SHIP in growth factor induced signalling,” *Prog Biophys Mol Biol*, vol. 71, no. 3-4, pp. 423–34, 1999.

[278] I. Tamir, J. C. Stolpa, C. D. Helgason, K. Nakamura, P. Bruhns, M. Daeron, and J. C. Cambier, “The RasGAP-binding protein p62dok is a mediator of inhibitory FcgammaRIIB signals in B cells,” *Immunity*, vol. 12, no. 3, pp. 347–58, 2000.

[279] B. Stork, K. Neumann, I. Goldbeck, S. Alers, T. Kahne, M. Naumann, M. Engelke, and J. Wienands, “Subcellular localization of Grb2 by the adaptor protein Dok-3 restricts the intensity of Ca2+ signaling in B cells,” *EMBO J*, vol. 26, no. 4, pp. 1140–9, 2007.

[280] D. W. Huang, B. T. Sherman, and R. A. Lempicki, “Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists,” *Nucleic Acids Res*, vol. 37, no. 1, pp. 1–13, 2009.

[281] H. Ogata, S. Goto, K. Sato, W. Fujibuchi, H. Bono, and M. Kanehisa, “KEGG: Kyoto Encyclopedia of Genes and Genomes,” *Nucleic Acids Res*, vol. 27, no. 1, pp. 29–34, 1999.

[282] A. Franceschini, D. Szklarczyk, S. Frankild, M. Kuhn, M. Simonovic, A. Roth, J. Lin, P. Minguez, P. Bork, C. von Mering, and L. J. Jensen, "STRING v9.1: protein-protein interaction networks, with increased coverage and integration," *Nucleic Acids Res*, vol. 41, no. Database issue, pp. D808–15, 2013.

[283] B. Brannetti, A. Via, G. Cestra, G. Cesareni, and M. Helmer-Citterich, "SH3-SPOT: an algorithm to predict preferred ligands to different members of the SH3 gene family," *J Mol Biol*, vol. 298, no. 2, pp. 313–28, 2000.

[284] T. Hou, W. Zhang, D. A. Case, and W. Wang, "Characterization of domain-peptide interaction interface: a case study on the amphiphysin-1 SH3 domain," *J Mol Biol*, vol. 376, no. 4, pp. 1201–14, 2008.