Net-PPI: Mapping the Human Interactome with Machine Learned Models

by

Kfir Schreiber

B.Sc., The Open University of Israel (2008)

Submitted to the Program in Media Arts and Sciences
in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2018

Author .... Signature redacted
Program in Media Arts and Sciences
August 10, 2018

Certified by..................... Signature redacted ..
Joseph M. Jacobson
Associate Professor of Media Arts and Sciences
Thesis Supervisor

Accepted by................. Signature redacted ..
Tod Machover
Academic Head, Program in Media Arts and Sciences

Net-PPI: Mapping the Human Interactome with Machine Learned Models

by

Kfir Schreiber

Submitted to the Program in Media Arts and Sciences
on August 10, 2018, in partial fulfillment of the
requirements for the degree of
MASTER OF SCIENCE

Abstract

The miracle of life is only possible thanks to a wide range of biochemical interactions between assortments of molecular agents. Amidst these agents, which enable all cellular activities, proteins are undoubtedly among the most important groups. Proteins facilitate countless intra- and inter-cellular functions, from regulation of gene expression to immune responses to muscle contraction, but they rarely act in isolation. These are the interactions between proteins, known as protein-protein interactions or PPIs, which sustain the fundamental role of proteins in all living organisms.

PPIs are also central to the study of diseases and development of therapeutics. Aberrant human PPIs are the primary cause of many life-threatening conditions, such as Alzheimer, Creutzfeldt-Jakob, and cancer; making the regulation of PPI activities a promising direction for pharmaceutical development. Despite the indisputable importance of PPIs, so far only a tiny fraction of all human PPIs has been discovered, and our current understanding of the core mechanisms and primary functionalities is insufficient.

While computational methods in general and machine learning in particular showed encouraging potential to address this challenge, their application in real-life has been limited. To mitigate this gap and make sure computational results perform as well in real-life, we introduce a set of gold-standard machine learning practices called NetPPI. The contributions of this thesis include NetPPI, a minimally-biased, carefully curated dataset of experimentally detected PPIs for training and evaluation of machine learning models; a comprehensive study of protein sequence representations for use with discriminative models; and data splitting methodology for machine learning purposes. We also present the Bilinear PPI model for state-of-the-art PPI prediction. Finally, we propose fundamental biological insight on the nature of PPIs, based on performance analysis of different prediction models.

Thesis Supervisor: Joseph M. Jacobson
Title: Associate Professor of Media Arts and Sciences

Net-PPI: Mapping the Human Interactome with Machine Learned Models

by

Kfir Schreiber

Submitted to the Program in Media Arts and Sciences
on August 10, 2018, in partial fulfillment of the
requirements for the degree of
MASTER OF SCIENCE
at the
Massachusetts Institute of Technology

The following people served as readers for the thesis:

Academic Adviser.................... Signature redacted

Joseph M. Jacobson, PhD

Associate Professor of Media Arts and Sciences, MIT

Reader .... Signature redacted ....

Kevin Esvelt, PhD

Assistant Professor of Media Arts and Sciences, MIT

Reader .... Signature redacted .

Aditya Khosla, PhD

Visiting Scientist, MIT

Co-Founder and Chief Technology Officer, PathAI

Acknowledgments

This amazing journey would not have been possible without the help of many people.

First and foremost, I would like to thank my advisor and mentor, Joe Jacobson. He has created a unique environment where people are encouraged to ask random questions, go after the greatest problems, and are given the perfect balance of guidance and independence.

I would like to thank my incredible lab-mates and friends. To Thras, for all the adventures we have and will face together. This wouldn't have been possible without him paving the way. To Noah, Lisa, Pranam, James, and Maksym for the conversations, brainstorming, and late-night ping-pong. They have made my days much happier.

I would like to give my gratitude to my readers, Kevin and Aditya, for the mentorship throughout this work. Their willingness to ask the hard questions was invaluable.

I would like to thank the great people of CBA, Neil, Shuguang, Fillipos, Prashant, Sam, Will, Ben, Amanda, Grace, Rui, Jake, and Amira. They are the most talented, driven, and inspirational people I have met. To Joe M., Ryan, and James for running the show in the best possible way.

My special thanks go to Linda, Keira, and the entire MAS staff, for their unparalleled devotion.

Last but not least, to my beloved family. My dear Ella, for her unconditional love and support, even when unconditional means leaving everything behind and moving to the other side of the world. To Mom and Dad, for giving me the courage to always pursue my dreams, knowing that their love is always waiting back at home. To Sapir and Omer, for being a constant inspiration.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

Proteins are the molecular machines that enable life. Using recombination of a minimal set of only 20 amino-acids, proteins manage to perform a vast diversity of biological functions, from structural functions to intercellular message delivery to immunological activities. However, proteins do not operate in isolation; rather, it is their tendency to form biochemical complexes with other proteins and various molecular agents that enables the functional diversity seen in living organisms. Proteins are involved in cellular pathways, working together with sugars, nucleotides, metals, fatty acids, and other proteins to form a network of biological pathways. Key to many pathways are interactions between two or more proteins, known as protein-protein interactions or PPIs. Among the primary roles of PPIs are cell-cycle control, immune reactions, metabolic catalyzation, transportation, and signal transduction. Protein-protein interactions are also a fundamental interest for the study of disease mechanisms and the development of therapeutics. Flawed PPIs, often caused by dysfunctional allosteric changes in one binding partner, are among the common triggers for a wide range of diseases, such as various types of cancer [1, 2, 3], Alzheimer's Disease [4], Creutzfeldt-Jakob Disease [5], and Huntington's Disease [6]. Despite the undeniable importance of PPIs and the vast research in the field, only a tiny fraction of the hypothesized interactions are known.

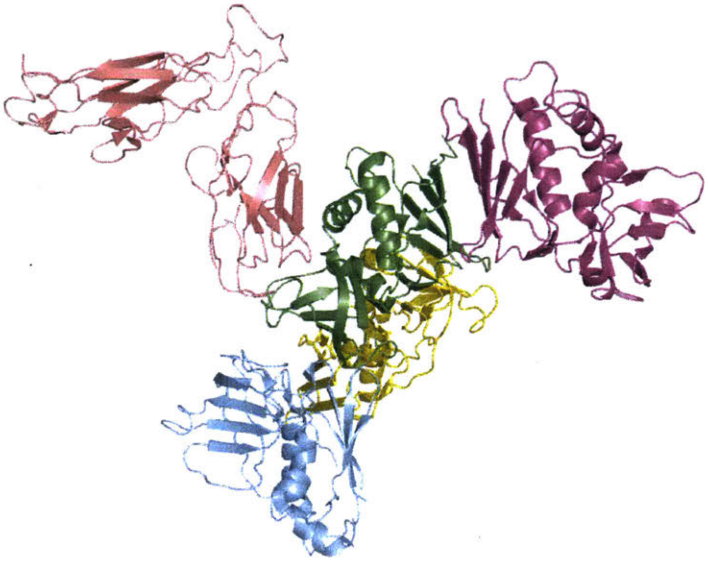


Figure 1-1: Protein complex of Streptococcal Pyrogenic Enterotoxin C (SpeC) with a human T cell receptor beta chain. Structure downloaded from RCSB (1KTK), visualized with PyMOL, and colored by chain.

Literature references to PPIs can be classified into two classes: physical PPIs describe associations between two or more proteins that bind together to a single macromolecular structure, while functional PPIs refer to proteins which are involved in the same functional pathway and do not necessarily interact physically. From this point forward we use the term PPI to describe physical binding or association between proteins. PPIs also come in various forms and shapes; proteins can interact with similar partners to form homodimers or with different partners to form heterodimers; they can also interact in pairs and multi-protein complexes involving more than two proteins. Overall, the PPI problem lies in the intersection of three biological fields. The Systems Biology approach aspires to identify existing interactions and map the complex Interactome network. Structural Biology is mostly looking to name the set of amino-acids, within a known pair of interacting proteins, which are critical for the interaction (also known as hot spots). Finally, Evolutionary Biology is interested in the fundamental question of how nature created such a diverse set of PPI functionalities under the strict cellular constraints and using only a minimal set of amino-acid building blocks. These different aspects for the same biophysical phenomena raise unique challenges and call for different paths to a solution.

For the reasons mentioned above, developing therapeutics to inhibit or activate PPIs is a long sought-after goal of the pharmaceutical industry. To date, the vast majority of PPI therapeutics are biologic compounds (e.g., antibodies, peptides), while small-molecule therapeutics usually fail to perform the required function. Small molecules, however, have several key advantages that make them extremely valuable. While, in most cases, biologics require intravenous administration, small molecules can be administered orally, thanks to their improved delivery mechanisms. Furthermore, small molecules can target intra-cellular therapeutic targets, an impossibility for most biologics. Therefore, small-molecule PPI inhibitors and activators have immense potential to target many "undruggable" targets and provide a cure to multiple diseases. However, development of small-molecule PPI therapeutics would have to address a non-trivial proportions challenge. Protein-protein interactions take place in large protein surfaces that might contain tens of residues. In most cases, even if a small-molecule binder existed, it would only block a limited segment of the interaction surface, leaving enough contact points for the PPI to continue undisturbed. Hot spots might hold the key to a solution to this problem. We now know that interactions are mediated by a small subset of critical residues, while the rest of the surface is insignificant. A small-molecule competitive binder that will block one or more of these critical residues will disturb the entire PPI (the same is true in the case of activators). Developing such therapeutics will require an incredibly detailed understanding of the PPI. Chapter 5 will discuss possible directions for such characterization of PPIs.

1.2 Related Work

1.2.1 Experimental Methods

Throughout the years, many experimental methods were developed to detect and characterize protein-protein interactions. Biophysical methods like X-ray crystallography, NMR spectroscopy, and atomic force microscopy rely on structural information to characterize PPIs. While biophysical methods provide comprehensive information about PPIs (e.g., binding mode, interaction dynamics) and are among the most accurate methods, they are also extremely resource- and time-consuming. Applications of biophysical methods are usually limited to a single or a few complexes at a time.

More recent in-vitro and in-vivo approaches aim to provide better throughput. Yeast two-hybrid (Y2H) is a high-throughput in-vivo Fragment Complementation Assay (Figure 1-2) widely used for PPI detection [7]. Y2H uses the two separable and functionally essential domains of the GAL4 protein in the yeast Saccharomyces cerevisiae to identify protein pairs that can form complexes. Although being the most common experimental technique for PPI detection, Y2H is tedious, limited in scale, and shows a high false detection rate (FDR). In-vitro techniques such as affinity purification methods (e.g. GST-pulldown [8, 9], co-immunoprecipitation[10]) and mass spectrometry suffer from similar disadvantages.

Both Fragment Complementation Assays and Affinity Purification methods can only provide supporting evidence for the existence of interaction but are not informative about the binding mechanism or the critical residues (hot spots) within the interacting proteins. Alanine scans were developed to overcome that gap and detect the hot spots of known interactions [12]. In Alanine scans, individual amino-acids are selectively mutated to Alanine. By measuring the $\Delta\Delta G$ of each mutant complex we can determine which residues are critical to the interaction. Alanine scans provide useful information about the binding mode of the PPI, but at the same time are expensive and extremely time-consuming, thus do not scale for more than a few complexes.

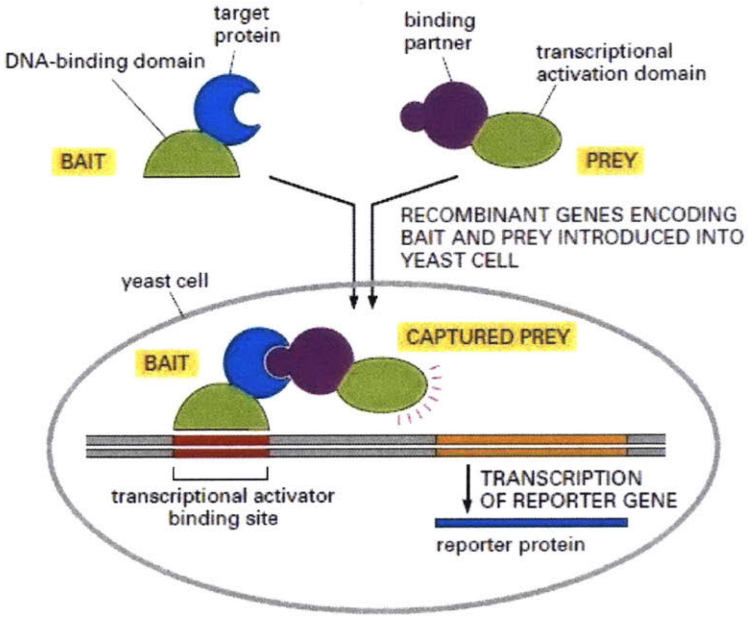

target protein
DNA-binding domain
binding partner
transcriptional activation domain
BAIT
PREY
RECOMBINANT GENES ENCODING BAIT AND PREY INTRODUCED INTO YEAST CELL
yeast cell
BAIT
CAPTURED PREY
transcriptional activator binding site
TRANSCRIPTION OF REPORTER GENE
reporter protein

Figure 1-2: Schematic description of a yeast two-hybrid system. **Source: [11]**

1.2.2 Computational Methods

As discussed in Section 1.2.1, experimental approaches enable small-scale characterization and bigger-scale detection of PPIs. Still, experiments can only provide limited coverage of the complete PPI landscape. Computational methods offer opportunities to overcome some of the limitations associated with their experimental alternatives, mostly due to their superior speed and lower cost. Structure-based computational methods like virtual docking algorithms [13, 14, 15] and computational Alanine scans [16] apply force fields or estimated score functions to approximate the free binding energy of a protein complex. This approach applies to both PPI detection and characterization. Although relatively accurate, these methods are computationally expensive and have highly limited applicability, due to the need for solved 3D structures for both the interacting partners.

Due to the limitations of experimental and structure-based computational methods, many researchers turned to sequence-based, structure-free computational approaches in the search for an interactome-scale detection algorithm. Recently, a large body of work has been generated, trying to predict PPIs directly from amino-acid sequences or sequence related features. Preliminary methods tried to identify binding sites in known proteins, independently of the interacting partners, by searching for known PPI motifs (i.e., known sequences of amino-acids, geometric patterns, and interacting domains) [17, 18], or by using the physiochemical properties of the amino acids in the sequence [19]. Succeeding works tried to predict the existence of an interaction by utilizing covariance in sequence mutations between interacting proteins [20], sequence homology [21], genomic context [22], and similarities in phylogenetic trees [23]. These approaches failed to deliver a complete and trusted mapping of the human interactome.

Data-Driven Methods

On a parallel path, in recent years, machine learning (ML) and especially deep neural networks (DNNs) have shown incredible results. While most ML concepts, like neural networks, pattern recognition, and backpropagation, trace back to the 1960's, it is the explosion in available data combined with the rapid growth in GPU-based computation that pushed learning algorithms to new horizons. Recent deep learning algorithms, a sub-field of machine learning, showed exceptional results in tasks that were once believed to be out of reach for machines, especially in pattern recognition tasks like image classification, natural language processing, and even audio generation. State-of-the-art image recognition algorithms identify objects in images with accuracy superior even to the human abilities [24]. Generative models create original text [25], audio [26], and images [27]. Reinforcement learning techniques are the engine behind machines with superhuman skills in the games of Go [28], Atari [29], and Dota2. Natural language processing was the core of the Jeopardy winner algorithm from IBM Watson [30]. Machine learning has had a significant impact on biology as well. Previous deep learning works were able to predict the sequence specificities of DNA- and RNA-binding proteins [31], the effects of non-coding variants in the human genome, and accurate 3D folds of small proteins [32].

The recent success of deep learning (DL) methods in various biological application raised high expectation for its performance in predicting PPIs. Bock and Gough were the first to apply a machine learning approach to the PPI problem, using a Support Vector Machine (SVM) model with the protein's primary sequence and associated physiochemical properties [33]. Subsequent works applied fusion of classifiers [34], hyperplanes [35], SVMs with Conjoint Triad (CT) representation [36], SVMs with auto-covariance [37], weighted sparse representation combined with discrete cosine transformation [38], random forests with multi-scale local feature representation [39], stacked auto-encoders [40], and a combination of a stacked sparse autoencoder with a probabilistic classification vector machine (PCVM) classifier [41].

However, impressive as they might be, reported results fail to imply relevance to real-life scenarios. Although prior works show constant improvement in prediction results, they all suffer from similar deficiencies, in the form of poor ML practices. These issues make the results unreliable and, in most cases, non-applicable in practice. The central issues that prevent reported results from genuine ability to generalize to utterly unseen proteins and interactions include the introduction of bias in the dataset construction phase; unfavorable splits into train, validation, and test sets; and the use of inadequate performance metrics.

Curation and construction of a negative dataset is a critical step in any ML algorithm and the first challenge in the PPI problem. Most of the existing PPI databases only report positive interactions, which leaves the authors with the task of estimating which protein pairs are unlikely to interact. A widely practiced approach uses subcellular localization as the deciding factor, under the very reasonable assumption that proteins in different subcellular compartments cannot interact. However, this approach introduces an undesired bias to the dataset, which results in predictions of protein co-localization rather than PPIs [42, 43, 44]. The next question to be answered is about splitting the data into training and validation sets. Many papers choose a random split, which means picking random pairs of proteins to be excluded from the training set and used as validation. However, protein sequences can be remarkably similar, which reduces the PPI prediction task to the much easier similarity scoring task. The algorithm only needs to memorize the training set and, at the validation-time, search for the most similar case in the memory. Some works try to avoid this issue by using a non-redundant dataset (i.e., a dataset where every protein sequence is unique above some threshold), but while this solves the similarity issue, the training and validation sets are still highly correlated since the same protein sequences are used in both (although with different pairings). An even more problematic approach [40] tries to strengthen the results by using external independent test sets composed of only positive samples. In all these case, when tested against an independent, positive and negative mixed, dataset, the prediction accuracy collapses to no more than a random guess. Lastly, some papers measure their algorithmic performance by inadequate metrics, like the use of binary accuracy with highly imbalanced datasets.

To address these issues, we follow the foot-steps of ImageNet [45] and MoleculeNet [46] and introduce NetPPI, a high-confidence, minimally-biased, and properly-split dataset of experimentally detected PPIs. Moreover, we propose a novel deep learning model architecture for prediction of PPIs, based on a fundamental hypothesis about the nature of PPI motifs.

1.3 Contributions

Chapter 2 of this thesis, introduces two versions of the NetPPI dataset: a non-redundant version, and an augmented version based on sequence homology. Chapter 3 surveys various possible representations for protein primary sequences, for usage with learning algorithms. Chapter 4 describes deep learning architectures for PPI detection, their relative performance, and the possible underlying structure of PPIs emerging from the results. Finally, Chapter 5 offers closing remarks and future research directions.

Chapter 2

Datasets

On two occasions I have been asked, "Pray, Mr. Babbage, if you put into the machine wrong figures, will the right answers come out?" ... I am not able rightly to apprehend the kind of confusion of ideas that could provoke such a question.

BABBAGE, CHARLES. PASSAGES FROM THE LIFE OF A PHILOSOPHER, 1864

Garbage in, garbage out

THE HAMMOND TIMES, 1957

2.1 Overview

Unsurprisingly, the first factor in the success of learned models is the data they rely on. This is also the first step in the design of machine learning algorithms. Carefully curated and widely accepted datasets can also set the foundations for rapid progress in a specific domain, as was seen in the fields of image recognition, after the publication of ImageNet, and chemoinformatics after the introduction of MoleculeNet. In addition to data curation, the rise of deep learning models should give a particular focus to the data splitting methods. Deep neural networks usually contain millions of tunable parameters, which allow them to display cutting-edge expressive abilities, but also, if not treated correctly, to memorize and overfit to the training set. Thus, inadequate data split into training and test sets might result in performance overestimation for models that do not have any real predictive power.

In this chapter, we will survey the existing PPI databases, mention past mistakes in the construction of datasets, and depict the methodology we use for building a high-confidence, minimally-biased dataset.

2.2 PPI Databases

The recent advancements in high-throughput experimental methods for PPI detection (e.g., yeast2hybrid) have resulted in a surge of PPI data and databases. At the time of writing (August 2018), there are 320 primary databases, meta-databases, and prediction databases available [47]. Primary databases curate experimentally detected PPIs from scientific literature; meta-databases syndicate several primary databases into a single resource; and prediction databases provide curated prediction from various experimental and computational methods. Databases also differ by organisms, level of curation, and level of confidence. Table 2.1 shows a breakdown of the largest PPI databases by size and type.

Name	Proteins	Interactions	Organism	Type	References
STRING	9,643,763	1,380,838,440	various	meta & prediction	[48], [49], [50], [51], [52], [53], [54], [55], [56], [57]
I2D	N/A	1,279,157	various	meta & prediction	[58], [59]
BioGRID	68,848	1,230,943	various	primary	[60]
IntAct	96,594	851,299	various	primary	[61]
APID	94,326	754,879	various	meta	[62]
InWeb_InBioMap	N/A	625,500	Human	meta	[63]
HIPPIE	N/A	325,468	Human	meta	[64], [65], [66], [67]
MINT	25,181	123,891	various	primary	[68]
HuRI	13,790	84,656	Human	primary	
DIP	28,826	81,762	various	primary	[69], [70], [71], [72]
HPRD	30,047	41,327	Human	primary	[73], [74], [75]

Table 2.1: PPI databases

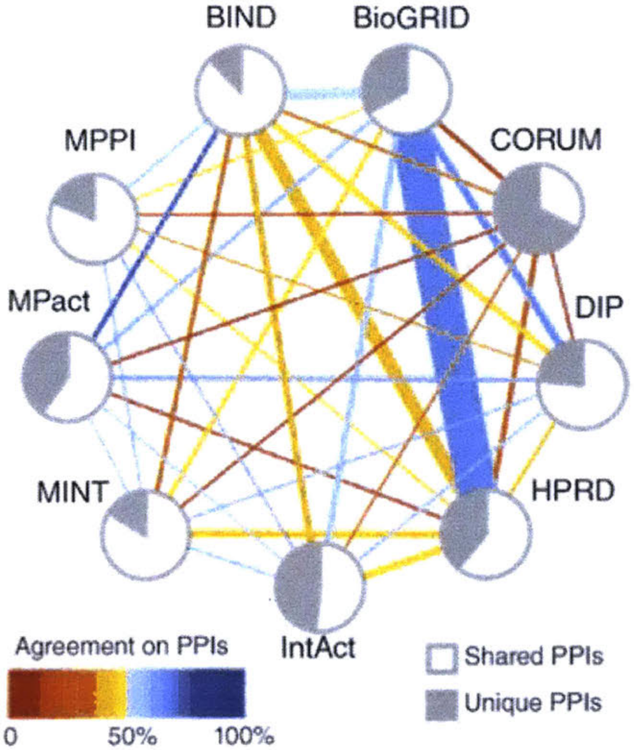


Figure 2-1: Curation overlap across different pairs of databases. Nodes represent individual databases, with the pie charts illustrating the proportion of shared and unique PPI records in each database. The edge thickness represents the number of instances where the two databases curate the same publication, whereas the edge color represents the average level of agreement (measured as defined above) on recorded interactions, following the color-coded scale. **Source: Turinsky et al. [76]**

While all primary databases curate PPIs from peer-reviewed scientific publications, the agreement among them is far from perfect. Turinsky et al. [76] studied the agreement levels between all major databases. Their results (shown in Figure 2-1) demonstrate the challenges in consolidating a single database for all PPIs. When outlining the dataset construction methodology, it is appropriate to also take into account the various types of PPI records that exist in all databases. Record types imply different relationships between the interacting partners, which roughly correlates to the level of confidence in the existence of a PPI. A full list of interaction types is shown in Table 2.2.

Type	IMEx ID	Definition
colocalization	MI:0403	Coincident occurrence of molecules in a given subcellular fraction observed with a low-resolution methodology from which a physical interaction among those molecules cannot be inferred.
Functional association	MI:2286	Binary relationship between biological entities when one of them modulates the other in terms of function, expression, degradation or stability of the other and the relationship between the partners cannot be ascertained as direct, so intermediate steps are implicitly present. This relation specifically does not imply a physical interaction between the entities involved.
genetic interaction	MI:0208	An effect in which two genetic perturbations, when combined, result in a phenotype that does not appear to be merely explained by the superimposition or addition of effects of the original perturbations.
Association	MI:0914	Interaction between molecules that may participate in the formation of one, but possibly more, physical complexes. Often describes a set of molecules that are co-purified in a single pull-down or co-immunoprecipitation but might participate in the formation of distinct physical complexes sharing a common bait.
Physical association	MI:0915	Interaction between molecules within the same physical complex. Often identified under conditions which suggest that the molecules are in close proximity but not necessarily in direct contact with each other.
direct interaction	MI:0407	Interaction between molecules that are in direct contact with each other.
covalent binding	MI:0195	Interaction leading to the formation of covalent bond within an autocatalytic molecule or between partners.

Table 2.2: PPI record types. IDs follow the naming set by The International Molecular Exchange Consortium (IMEx). Definitions were taken from The European Bioinformatics Institute (EMBL-EBI) website.

2.3 NetPPI - Curation Methodology

In this section, we describe the curation process behind NetPPI, a high-confidence bias-free dataset for training and validation of learning algorithms. The primary guideline throughout this section is "quality over quantity." For the first time in the history of PPI research, the rapid growth in the available experimental data allows us to include only very high-confidence data points in our dataset. This is especially important for deep learning methods, which are known to be data demanding but also highly susceptible to overfitting. Following this guideline, we do not use information from prediction databases and focus solely on experimental data. Meta-databases make it easier to obtain and prepare the data in a centralized way, however, they also introduce another error-prone step to the pipeline. Therefore, we collect data from primary databases only. We also filter out proteins that are shorter than 50 amino acids, since they are likely to be protein fragments.

Positive dataset. Positive PPIs are combined from five primary databases, INstruct, Database of Interacting Proteins (DIP), IntAct, The Molecular INTeraction Database (MINT), and BioGrid. Consecutively, the records are filtered by interaction type, so only direct interactions are included in the dataset. To guarantee that only high-confidence PPIs are included in the dataset, we accept a record if and only if it complies with at least one of the following conditions: (a) there exists a solved crystal structure of the protein complex, (b) the same PPI was discovered by two unique experimental methods, (c) the PPI is supported by two separate publications. Finally, duplicated records were removed. This resulted in a positive set of 240290 interactions across 48834 unique protein sequences.

Negative dataset. Negative samples are exactly as important for discriminative models as positive ones. Bias introduced during the construction of the negative dataset might result in a task that is much easier than intended and a model that is unable to generalize to real-life unseen samples. Therefore, careful attention is required when choosing the negative sampling methodology for the PPI problem. Prior publications suggested several approaches:

- **Co-localization**- protein pairs with different sub-cellular localization.

- **Functional annotations**- protein pairs with significantly different functional annotations.

- **Random sampling**- random sampling of unobserved interactions.
- **Unobserved experimental samples**- random sampling among unobserved pairs after experimental post-processing.

Co-localization and functional annotations were shown to bias the prediction towards sub-cellular localization and functional classification, respectively, instead of interaction prediction. Therefore, NetPPI takes a hybrid approach between random sampling and unobserved experimental samples. Negative experimental samples were taken from large-scale two-hybrid assays [77]. Trabuco et al. describe a method for extracting negative interactions from two-hybrid data, with a confidence score that is based on the length of the shortest path between the non-interacting proteins. We choose a minimum confidence score of 5, which results in 15568532 negative interactions among 17650 unique proteins. Since the experimental negatives only cover a small percentage of the proteins involved in positive interactions, we augment the experimental negatives with random sampling of additional 15568532 unobserved pairs of proteins (i.e., where each protein is involved in at least one positive interaction). According to Launay et al. [78] the expected false negative rate for our random sampling is 0.2%, which sets the overall false negative rate of our dataset at around 0.1%.

Species. The mission of this thesis, as well as other works, is to detect and characterize **Human PPIs**. However, for some algorithms, it might prove useful to include in the training set PPIs from other species as well. Therefore we make available two datasets: NetPPI-NR contains a non-redundant version of the interactions mentioned above, and NetPPI-HNR contain only the Human interactions described in the previous sections.

Data augmentation. Data augmentation is a common practice in machine learning to enrich a dataset with additional synthetic samples. In our case, synthetic samples can be generated by sequence homology. To find homologs, we use the Basic Local

Alignment Search Tool(BLAST) [79] provided by the National Center for Biotechnology Information (NCBI), and the UniProt database [80]. In the augmented dataset, all homologous proteins are assumed to facilitate the same interactions.

Overall we described four datasets:

- **NetPPI** - an augmented version including all species.
- **NetPPI-NR** - a non-augmented, non-redundant version including all species.
- **NetPPI-H** - a Human only, augmented version.
- **NetPPI-HNR** - a Human only, non-augmented, non-redundant version.

2.4 Train/Validation/Test Split

The last preparation step, and the goal of this section, is to split the dataset into training, validation and test subsets. A correctly performed split will enable a fair evaluation of prediction models and guarantee that the results indeed imply a true ability to generalize. We first define three generalization categories:

- **Novel proteins (C1)** - detection of interactions between two unseen proteins.
- **Novel partners (C2)** - detection of interactions between a protein from the training set and an unseen partner.
- **Novel pairings (C3)** - detection of interactions between two proteins from the training set, which were not seen as a pair.

For split purposes in all categories, we consider proteins to be different if and only if they share less than 25% sequence similarity. For example, in the C2 category, the unseen partner cannot share more than 25% sequence similarity with any protein in the training set. Sequence similarity was calculated using the *pairwise2* function from the BioPython package [81].

Chapter 3

Data Representation

3.1 Overview

The previous chapter discussed the key challenges and opportunities that occur when choosing which data to use for training and evaluating prediction models. An additional key parameter in the success of ML models is the data representation. The goal is to find a set of features that describe the data in a meaningful enough way for the algorithm to differentiate between positive and negative samples. By nature, protein sequences are discrete objects with variable length, and these are the main challenges any representation needs to address. The next sections describe the most common representations for protein sequences.

3.2 Representations

One-Hot (OH) Encoding is the gold standard for encoding categorical data (e.g., amino acid characters). OH encoding requires a specific ordering of the classes, which we will assume, without loss of generality, to be the alphabetical ordering of the amino acid characters. Then each character in the sequence is represented as a vector of binary values, where the only '1' value is in the position which corresponds to the position of the relevant amino acid in the order. Aggregation of all vectors results in a binary matrix of size $L * 20$, where L is the length of the sequence, and

$$OH(i,j) = \begin{cases} 1 & p_j = a_i \\ 0 & o.w. \end{cases} \tag{3.1}$$

where OH is the One-Hot matrix, $P = (p_1, p_2, ..., p_L)$ is the protein sequence, p_j is the j'th residue in the sequence, and a_i is the i'th amino-acid by the alphabetical ordering.

For example, the protein sequence $MMADRSIMARG$ will be encoded to the following One-Hot matrix,

$$\begin{bmatrix}
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0
\end{bmatrix}^T$$

Physiochemical Features. OH encoding treats residues as discrete classes with no implied relations between them. However, amino acids are physical objects with physiochemical properties, and as such they imply various similarity landscapes. An amino acid can be more similar to another than to a third. To utilize this additional information, the Physiochemical Features approach represents a residue as a vector of real values that correspond to a list of physiochemical properties. For example the same $MMADRSIMARG$ sequence will be encoded to the following matrix,

$$\begin{bmatrix}
0.64 & -1.3 & 0.002683 & 5.7 & 0.221 & 2.034 & 94.1 \\
0.64 & -1.3 & 0.002683 & 5.7 & 0.221 & 2.034 & 94.1 \\
0.62 & -0.5 & 0.007187 & 8.1 & 0.046 & 1.181 & 27.5 \\
-0.9 & 3 & -0.02382 & 13 & 0.105 & 1.587 & 40 \\
-2.53 & 3 & 0.043587 & 10.5 & 0.291 & 2.56 & 105 \\
-0.18 & 0.3 & 0.004627 & 9.2 & 0.062 & 1.298 & 29.3 \\
1.38 & -1.8 & 0.021631 & 5.2 & 0.186 & 1.81 & 93.5 \\
0.64 & -1.3 & 0.002683 & 5.7 & 0.221 & 2.034 & 94.1 \\
0.62 & -0.5 & 0.007187 & 8.1 & 0.046 & 1.181 & 27.5 \\
-2.53 & 3 & 0.043587 & 10.5 & 0.291 & 2.56 & 105 \\
0.48 & 0 & 0.179052 & 9 & 0 & 0.881 & 0
\end{bmatrix}^T$$

In this thesis we follow the work of Guo et al. [37] and use seven physiochemical features, hydrophobicity, hydrophilicity,net charge index of side chains, polarity, polarizability, solvent accessible surface area, and the volume of side chains. The physiochemical values used are shown in Table B.3.

Conjoint Triad (CT), first proposed by Shen et al. [36], also incorporates physiochemical information, but is doing so by clustering the 20 amino acids into 7 clusters by side chain volume and dipole (complete list is shown in Table B.4). After replacing each amino acid with the equivalent cluster, a sliding window of three amino acids (triads) is used to calculate the occurrence frequency for each of the 343 possible triads. In this case, our example sequence $MMADRSIMARG$ will be converted to 33165323151, and the feature vector will be 343-dimensional with all values equal to zero but the corresponding positions to the 331, 316, 165, 653,532, 323, 231, 315, 151 triads (which will be equal to 1/9).

Substitution Matrix Representation (SMR) adds evolutionary information to the sequence representation. In SMR, each amino acid is replaced by a 20-dimensional vector, where the i'th value represents the mutation probability between the original amino acid and a_i. This results in a matrix of shape $L * 20$. It is standard to use one of the BLOSUM matrices for this purpose, and here we choose the BLOSUM62 (Table B.5). With SMR, $MMADRSIMARG$ will become

$$\begin{bmatrix}
-1 & -1 & -2 & -3 & -1 & 0 & -2 & -3 & -2 & 1 & 2 & -1 & 5 & 0 & -2 & -1 & -1 & -1 & -1 & 1 & -3 & -1 & -1 & -4 \\
-1 & -1 & -2 & -3 & -1 & 0 & -2 & -3 & -2 & 1 & 2 & -1 & 5 & 0 & -2 & -1 & -1 & -1 & -1 & 1 & -3 & -1 & -1 & -4 \\
4 & -1 & -2 & -2 & 0 & -1 & -1 & 0 & -2 & -1 & -1 & -1 & -1 & -2 & -1 & 1 & 0 & -3 & -2 & 0 & -2 & -1 & 0 & -4 \\
-2 & -2 & 1 & 6 & -3 & 0 & 2 & -1 & -1 & -3 & -4 & -1 & -3 & -3 & -1 & 0 & -1 & -4 & -3 & -3 & 4 & 1 & -1 & -4 \\
-1 & 5 & 0 & -2 & -3 & 1 & 0 & -2 & 0 & -3 & -2 & 2 & -1 & -3 & -2 & -1 & -1 & -3 & -2 & -3 & -1 & 0 & -1 & -4 \\
1 & -1 & 1 & 0 & -1 & 0 & 0 & 0 & -1 & -2 & -2 & 0 & -1 & -2 & -1 & 4 & 1 & -3 & -2 & -2 & 0 & 0 & 0 & -4 \\
-1 & -3 & -3 & -3 & -1 & -3 & -3 & -4 & -3 & 4 & 2 & -3 & 1 & 0 & -3 & -2 & -1 & -3 & -1 & 3 & -3 & -3 & -1 & -4 \\
-1 & -1 & -2 & -3 & -1 & 0 & -2 & -3 & -2 & 1 & 2 & -1 & 5 & 0 & -2 & -1 & -1 & -1 & -1 & 1 & -3 & -1 & -1 & -4 \\
4 & -1 & -2 & -2 & 0 & -1 & -1 & 0 & -2 & -1 & -1 & -1 & -1 & -2 & -1 & 1 & 0 & -3 & -2 & 0 & -2 & -1 & 0 & -4 \\
-1 & 5 & 0 & -2 & -3 & 1 & 0 & -2 & 0 & -3 & -2 & 2 & -1 & -3 & -2 & -1 & -1 & -3 & -2 & -3 & -1 & 0 & -1 & -4 \\
0 & -2 & 0 & -1 & -3 & -2 & -2 & 6 & -2 & -4 & -4 & -2 & -3 & -3 & -2 & 0 & -2 & -2 & -3 & -3 & -1 & -2 & -1 & -4
\end{bmatrix}^T$$

Position Specific Scoring Matrix (PSSM) is another evolution-based method, but while SMR uses mutation probabilities across all existing sequences, PSSM only considers homologos of the sequence of interest. In a similar way the final representation is a matrix of size $L * 20$, where high PSSM scores correlate to highly preserved parts of the protein sequence. In this work, we use the PSI-BLAST package (parameters specified in Section B.1).

3.3 Handling Varying Lengths

So far, other than Conjoint Triad, all the representations resulted in matrices of varying lengths. However, most ML algorithms require fixed sized inputs. This section describes two methods for processing sequences of different length into fixed sized matrices.

Zero Padding. The sequence or the relevant feature matrix are padded with zeros at the end until a specific maximal length. For example, using the $MADR$ sequence with the OH Encoding representation and a maximal length of 7 will result in

$$\begin{bmatrix}
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0
\end{bmatrix}^T$$

Auto-Covariance (AC) takes a different approach. Instead of adding values to short sequences, AC compresses variable length sequences into a fixed size matrix. It is done by only considering the correlations between residues with a specific distance between them, up to a maximal distance D. AC can be applied to any numeric representation but is most commonly used with physiochemical features [82, 83, 40]. First, the numeric features are normalized by the following equations,

$$X(i,k) = \frac{A(i,k) - A_{min}(k)}{A_{max}(k) - A_{min}(k)} \tag{3.2}$$

$$\tilde{X}(i,k) = \frac{A(i,k)}{\sqrt{\sum_{j=1}^{20} X^2(j,k)}} , \tag{3.3}$$

where A is the feature matrix, $A_{min}(k)$ and $A_{max}(k)$ are the minimum and maximum values of the k'th feature, respectively, and $\tilde{X}$ is the normalized feature matrix. Then AC is defined as

$$AC(d,k) = \frac{1}{L-d}\sum_{i=1}^{L-d}(\tilde{X}(i,k) - \frac{1}{L}\sum_{j=1}^{L}\tilde{X}(j,k))(\tilde{X}(i+d,k) - \frac{1}{L}\sum_{j=1}^{L}\tilde{X}(j,k)) , \tag{3.4}$$

where AC is the auto-covariance matrix (of shape $D*K$), L is the sequence length, and d is the distance along the sequence between a residue and its neighbor. The hyperparameter D, the maximal distance, was studied by Guo et al.[37], who suggested that $D = 30$ is the optimal value. However their assessment was limited to the specific model of choice (SVM) and the specific validation set. In this work, we use the suggested value, as well as the maximal possible value of $D = 50$ (i.e., the length of of the shortest sequence).

Chapter 4

Uncovering the Nature of PPIs through Predictive Modeling

4.1 Overview

Previous chapters outlined a set of gold-standard machine learning practices for the PPI problem. This chapter will attempt to bring together prediction algorithms, built on top of the standards described in Chapters 2 and 3, and the fundamental question about the underlying structure of PPIs. The core hypothesis is that *sequence motifs* are re-used by nature across proteomes to design new PPIs. Here we use the term 'motifs' under its widest definition, and in the next sections, we will show that careful analysis of the prediction results might shed some light on the real nature of those PPI motifs.

4.2 Baseline Models

We start by defining a set of baseline models, which will be used to evaluate the strengths and weaknesses of our architectures. They will also provide intuition about how much of the predictive power of our algorithms can be attributed to simple similarities between datasets and representations.

K-Nearest Neighbor (KNN)

One of the simplest machine learning classifiers, KNN is a non-parametric learning algorithm that implies a lazy learning approach, which means the training step involves simply memorizing the training data, and the computation is postponed to the inference step. In the inference step, the algorithm searches for the K closest (by some distance metric) training samples to the query sample. Subsequently, the K samples "vote" according to their labels and a weighting schema. Here we use the Euclidean distance and weight samples by the inverse of their distance. Moreover, we optimize the hyper-parameter K through a random search and use the physiochemical AC representation. To accelerate training and improve predictions, we used a Bagging approach [84] and trained 32 classifiers, where each was trained on 10% of the available data. We also applied down-sampling to deal with the imbalance between positive and negative samples. The scikit-learn package [85] was used for implementation.

Support Vector Machines (SVM)

SVM is a supervised discriminative classifier used in many machine learning applications. During training, a linear SVM algorithm tries to construct single or multiple hyperplanes that will separate samples from different classes. Non-linear SVMs use the "kernel trick" to first map the samples to a much higher dimensional space and then build the separating hyperplanes. The optimal separating hyperplanes are the

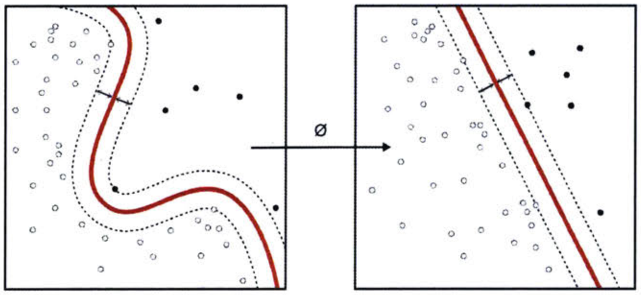


Figure 4-1: Non-linear SVM classifier. **Image credit: Alisneaky - Own work, CC0, https://commons.wikimedia.org/w/index.php?curid=14941564**

ones that achieve the maximal distance to the nearest training sample from each class (Figure 4-1). Here we use non-linear SVM with a Gaussian Radial Basis Function kernel. Two representations were tested, Auto-Covariance with physiochemical features and Auto-Covariance with physiochemical features and SMR. Like in the KNN case, we used a Bagging classifier with 32 estimators and down-sampling of negative samples. The C hyperparameters were chosen by cross-validation. Again, scikit-learn was used.

Stacked Autoencoder

The final baseline model is the current state-of-the-art stacked autoencoder with physiochemical AC representation developed by Sun et al. [40]. To make sure we fairly compare the performance, a new model was trained, using the reported architecture and hyper-parameters, with the NetPPI datasets.

4.3 Model Designs

To explore our hypothesis about the structure of PPIs, we introduce the Bilinear-PPI (B-PPI) model. B-PPIs are Deep Neural Networks (DNNs) following the general schema outlined by Lin et al. [86] in their work on Bilinear-CNN (B-CNN) models, which has shown state-of-the-art performance in fine-grained visual recognition tasks. At the core of the proposed architecture are two DNNs that operate as feature extractors, each on a different protein sequence. The feature extractors are then followed by a bilinear aggregation operation, and several fully connected layers (Figure 4-2). We experiment with several DNN feature extractors, as well as two bilinear aggregation operations, the cross-product, and concatenation. In case PPIs are indeed recombinations of re-used motifs and assuming a feature extractor capable of identifying them, we would expect B-PPI to exhibit superior performance compared to the baseline models. Next, we discuss several DNN feature extractors.

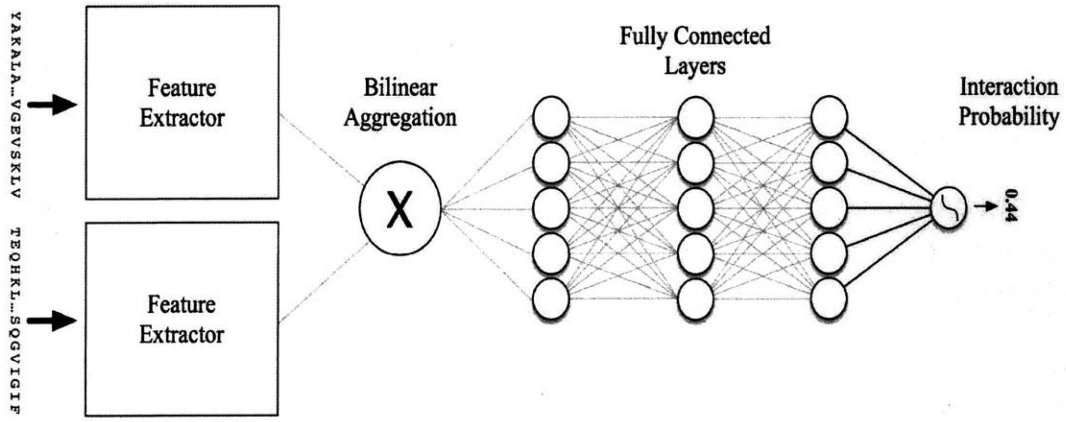


Figure 4-2: Template architecture for Bilinear-PPI networks.

Convolutional Motif Extraction Tool (CoMET)

CoMET, developed by Karydis et al. [87], is a data-driven computational tool for hierarchical decomposition of protein sequences into motifs of arbitrary length. At its core, CoMET is a Deep Convolutional Neural Network (ConvNet) in an encoder architecture. It was shown that CoMET can recognize known and novel **contiguous** sequence motifs. Here we use the CoMET architecture to assess the hypothesis that PPIs are mediated mostly through contiguous preserved sequence motifs. Key to the CoMET profile is the global Max Pooling (gMP) layer that follows the convolutional layers. The gMP guarantees that each motif detector (i.e., convolutional filter) is activated by one and only one contiguous part of the protein sequence. The original work suggests using fully-connected layers after the gMP for various prediction tasks. However, when even a single fully-connected layer is used, we lose the guarantee that the model extracts contiguous motifs. Therefore, in this model, we view the output of the gMP as the extracted motifs. The numbers of motif detectors and convolutional layers are optimized as hyper-parameters.

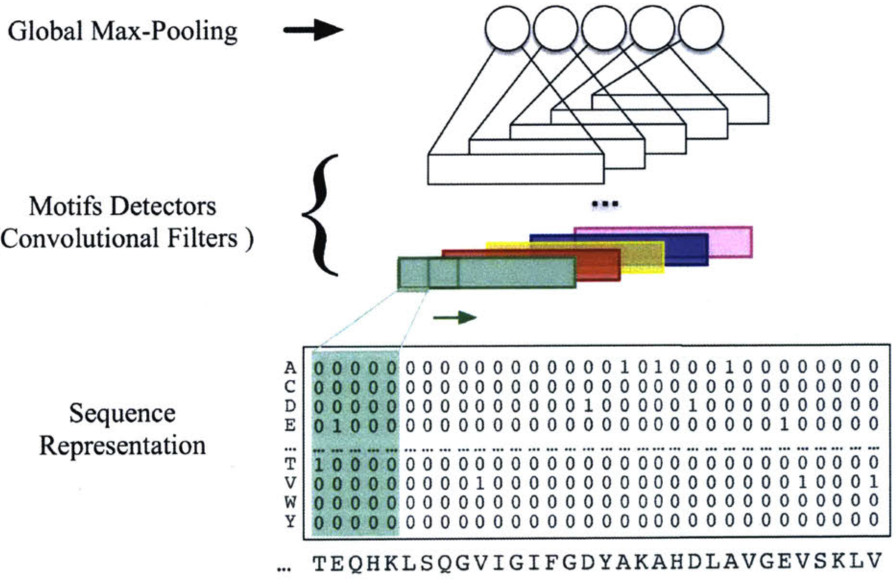
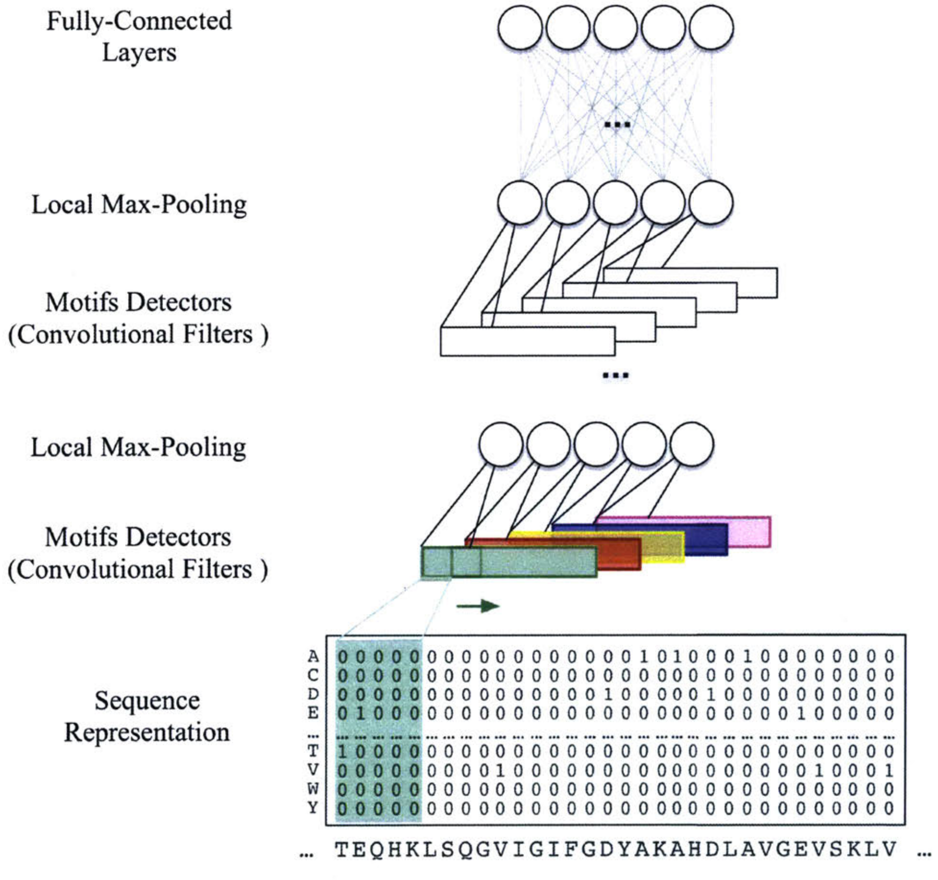

Global Max-Pooling
Motifs Detectors
(Convolutional Filters)
...
Sequence
Representation
A 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 1 0 0 0 1 0 0 0 0 0 0 0 0
C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
D 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0
E 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0
...
T 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
V 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 1
W 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
Y 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
... TEQHKLSQGVIGIFGDYAKAHDLAVGEVSKLV ...

Figure 4-3: Network architecture for the CoMET feature extractor.

Non-Contiguous CoMET (NC-CoMET)

NC-CoMET is an adaptation of the original CoMET architecture, designed to allow for non-contiguous motifs. First, the global Max Pooling is replaced with a **local** Max Pooling of size 2 after each convolutional layer. This change allows for each motif detector to be partially activated by various parts of the sequence. Additionally, we apply a set of fully-connected layers to the output of the final local Max Pooling. Again, the numbers of motif detectors and convolutional and fully-connected layers were optimized as hyper-parameters.



```
A 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 1 0 0 0 1 0 0 0 0 0 0 0 0
C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
D 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0
E 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0
... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ...
T 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
V 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 1
W 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
Y 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
... T E Q H K L S Q G V I G I F G D Y A K A H D L A V G E V S K L V ...
```

Figure 4-4: Network architecture for the NC-CoMET feature extractor.

Densely Connected Convolutional Networks (DenseNet)

DenseNet is a state-of-the-art neural network for visual recognition tasks [88]. DenseNets use inter-layer connections between the convolutional layers, to support substantially deeper convolutional networks. This approach has shown superior expressive capabilities compared to conventional ConvNets. By choosing to use DenseNets, we hope to minimize the likelihood of an inadequate feature extractor being the bottleneck for the model performance.

Hyper-parameter optimization for all models was done using the Bayesian Hyperband method described in Appendix A.

4.4 Evaluation Metrics

Performance evaluation will be conducted using the following metrics:

- Balanced Accuracy $= \frac{1}{2}(\frac{TP}{TP+FN} + \frac{TN}{TN+FP})$
- Specificity $= \frac{TN}{TN+FP}$
- Sensitivity (Recall) $= \frac{TP}{TP+FN}$
- Precision $= \frac{TP}{TP+FP}$
- F1 Score $= \frac{2TP}{2TP+FP+FN}$
- Area Under the Receiver Operating Characteristic Curve (ROC AUC or simply AUC)
- Matthews correlation coefficient (MCC)

where TP, TN, FP, and FN are true positive, true negative, false positive, and false negative, respectively.

4.5 Results

As expected, the results from the KNN and SVM baseline models (Table 4.1) demonstrate a clear difference between categories of generalization. The KNN model achieved balanced accuracy scores of 0.546, 0.695, and 0.89; AUC scores of 0.562, 0.768, and 0.956; and MCC scores of 0.025, 0.315, and 0.75 on the C1, C2, and C3 categories, respectively. The same trend was observed for the SVM models as well. The fact the simple machine learning models perform so well on the C3 category proves our assumption that generalization to unseen pairings alone is a straightforward task, and the random split an inadequate method to evaluate predictive models. Therefore the next sections will focus on the more challenging C2 and C1 categories.

Model Type	Representation	Category	Balanced Accuracy	Specificity	Sensitivity (Recall)	Precision	F1 Score	AUC	MCC
KNN	Physiochemical Auto-Covariance	C1	0.546	0.385	0.707	0.020	0.039	0.562	0.025
		C2	0.695	0.637	0.753	0.342	0.47	0.768	0.315
		C3	**0.89**	**0.932**	**0.848**	**0.759**	**0.801**	**0.956**	**0.75**
SVM	Physiochemical Auto-Covariance	C1	0.506	0.021	**0.992**	0.017	0.035	0.521	0.012
		C2	0.681	0.407	0.956	0.021	0.041	0.701	0.084
		C3	**0.873**	**0.812**	0.934	**0.554**	**0.696**	**0.96**	**0.632**
SVM	Physiochemical Auto-Covariance with SMR	C1	0.556	0.287	0.825	0.020	0.039	0.622	0.032
		C2	0.693	0.565	0.823	0.321	0.462	0.754	0.31
		C3	**0.852**	**0.86**	**0.844**	**0.601**	**0.702**	**0.917**	**0.626**

Table 4.1: Performance of baseline models across generalization categories. Categories: C1 - novel proteins, C2 - novel partners, C3 - novel pairings. Results in bold represent the best category per model.

C2 - Generalizing to Unseen Partners

Generalization to novel partners (i.e., the C2 category) poses a bigger predictive challenge and can be of interest in some real-life applications. This section (Table 4.2) evaluates the ability of our model to identify interactions between proteins from the training set and proteins outside of it. This can be used to map all interactions of proteins with partially known interactions. In this category, our NC-CoMET model with the Physiochemical Auto-Covariance with SMR and physiochemical and SMR features with zero-padding representations display the best performance and outperform the previous state-of-the-art.

Model Type	Representation	Balanced Accuracy	Specificity	Sensitivity (Recall)	Precision	F1 Score	AUC	MCC
KNN	PC-AC	0.695	0.637	0.753	0.342	0.47	0.768	0.315
SVM	PC-AC	0.681	0.407	**0.956**	0.021	0.041	0.701	0.084
	PC-SMR-AC	0.693	0.565	0.823	0.321	0.462	0.754	0.31
Stacked Autoencoder	PC-AC	0.575	**0.992**	0.159	0.179	0.168	0.733	0.16
CoMET	PC-AC	0.701	0.654	0.749	0.309	0.438	0.769	0.308
	PC-SMR-AC	0.694	0.753	0.635	0.347	0.449	0.764	0.315
	PC-ZP	0.648	0.613	0.683	0.267	0.384	0.709	0.225
	PC-SMR-ZP	0.687	0.692	0.681	0.314	0.43	0.754	0.291
	OH-ZP	0.687	0.708	0.665	0.321	0.433	0.754	0.294
	OH-SMR-ZP	0.73	0.685	0.775	0.338	0.47	0.8	0.355
NC-CoMET	PC-AC	**0.741**	0.845	0.763	**0.36**	**0.489**	0.818	**0.377**
	PC-SMR-AC	0.677	0.725	0.63	0.025	0.047	0.75	0.082
	PC-ZP	0.725	0.573	0.877	0.298	0.445	0.825	0.339
	PC-SMR-ZP	0.638	0.33	0.947	0.226	0.366	**0.835**	0.232
	OH-ZP	0.645	0.35	0.94	0.23	0.37	0.827	0.239
	OH-SMR-ZP	0.674	0.459	0.889	0.254	0.395	0.788	0.268

Table 4.2: Performance on the C2 category. Representations: PC-AC - Physiochemical Auto-Covariance, PC-SMR-AC - Physiochemical Auto-Covariance with SMR, PC-ZP - Physiochemical features with zero-padding, PC-SMR-ZP - Physiochemical and SMR features with zero-padding, OH-ZP - One-hot encoding with zero-padding, OH-SMR-ZP - One-hot encoding and SMR features with zero-padding.

C1 - Generalizing to Unseen Proteins

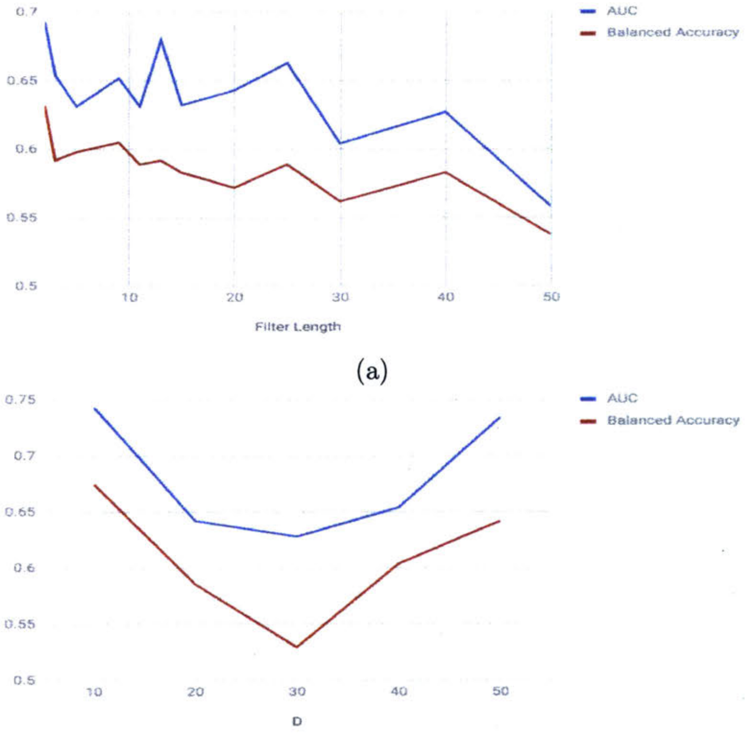
The greatest goal of a PPI detector is to detect interactions between completely unseen proteins. It is also the most informative category about the nature of PPI. In this category, the baseline models (i.e., KNN and SVMs) achieve a balanced-accuracy of $53\,\% \pm 3\,\%$, which suggests that simple similarities are not enough for the required generalization and that any improvement must indicate on the existence of an underlying structure in the data. The performance of the different models is shown in Table 4.3. Again, our models outperform the state-of-the-art Stacked Autoencoder, and the CoMET model with the Physiochemical Auto-Covariance with SMR representation displays the optimal performance.

Model Type	Representation	Balanced Accuracy	Specificity	Sensitivity (Recall)	Precision	F1 Score	AUC	MCC
KNN	PC-AC	0.546	0.385	0.707	0.020	0.039	0.562	0.025
SVM	PC-AC	0.506	0.021	**0.992**	0.017	0.035	0.521	0.012
	PC-SMR-AC	0.556	0.287	0.825	0.020	0.039	0.622	0.032
Stacked Autoencoder	PC-AC	0.636	0.994	0.277	0.433	0.338	0.718	**0.339**
CoMET	PC-AC	0.635	0.599	0.671	0.267	0.383	0.7	0.208
	PC-SMR-AC	**0.675**	0.687	0664	0.316	**0.429**	**0.743**	0.278
	PC-ZP	0.596	0.553	0.638	0.238	0.346	0.643	0.147
	PC-SMR-ZP	0.618	0.553	0.683	0.245	0.366	0.679	0.181
	OH-ZP	0.55	0.33	0.77	0.2	0.318	0.594	0.082
	OH-SMR-ZP	0.594	0.322	0.867	0.218	0.348	0.696	0.16
NC-CoMET	PC-AC	0.577	0.416	0.739	0.216	0.335	0.633	0.122
	PC-SMR-AC	0.632	0.536	0.728	0.255	0.378	0.713	0.202
	PC-ZP	0.603	0.398	0.808	0.226	0.354	0.655	0.164
	PC-SMR-ZP	0.617	0.655	0.58	0.268	0.366	0.664	0.184
	OH-ZP	0.544	0.203	0.885	0.017	0.034	0.613	0.027
	OH-SMR-ZP	0.518	0.199	0838	0.185	0.304	0.573	0.036
DenseNet	PC-AC	0.507	**0.998**	0.016	**0.638**	0.032	0.507	0.082
	PC-SMR-AC	0.494	0.977	0.012	0.097	0.021	0.494	-0.031
	PC-ZP	0.522	0.401	0.642	0.19	0.293	0.529	0.035
	PC-SMR-ZP	0.517	0.329	0.706	0.187	0.295	0.533	0.028
	OH-ZP	0.52	0.495	0.545	0.191	0.282	0.528	0.03
	OH-SMR-ZP	0.516	0.191	0.842	0.185	0.303	0.553	0.032

Table 4.3: Performance on the C1 category. Representations: PC-AC - Physiochemical Auto-Covariance, PC-SMR-AC - Physiochemical Auto-Covariance with SMR, PC-ZP - Physiochemical features with zero-padding, PC-SMR-ZP - Physiochemical and SMR features with zero-padding, OH-ZP - One-hot encoding with zero-padding, OH-SMR-ZP - One-hot encoding and SMR features with zero-padding.

Motif Length

In order to better understand the structure of PPI motifs, we explored different configurations of the CoMET architecture. The goal was to map the impact of the motif detector size on the overall performance. To isolate other effects, we used a basic architecture with just one convolutional layer, 100 motif detectors (i.e., filters), and 3 fully-connected layers with 400 units each after the bilinear aggregation. We experimented with two representations: the physiochemical and SMR features with one-hot encoding, and the physiochemical and SMR features with Auto-Covariance. We tested different filter lengths and different D values, respectively. The results are shown in Figure 4-5.



(a)

(b)

Figure 4-5: Impact of motif detector size on performance. (a) AUC and balanced accuracy scores for different filter lengths, using the physiochemical and SMR features with one-hot encoding representation. (b) AUC and balanced accuracy scores for different D values, using the physiochemical and SMR features with Auto-Covariance representation.

4.6 Discussion

In this section, we first demonstrated the critical role of a proper data split in designing and evaluating predictive models. The gold-standard machine learning practices for PPI prediction described in Chapter 2 guarantee that the trained models perform in real-life just as well as in training. To the best of our knowledge, this work is the first to apply rigorous analysis and appropriate validation metrics to the task of predicting PPIs directly from sequence. The results of the described models on the C2 and especially the C1 categories suggest that nature is indeed reusing some PPI motifs to build new interactions.

Further investigation might still be required to asses exactly how much of the interactome can be explained with the known motifs, but we can use the results described in this work to calculate a lower bound. The expected balanced accuracy for any model on proteins with no PPI motif is 0.5; therefore, any improvement can be used to bound the number of proteins which contain at least one PPI motif. Considering the CoMET OH-SMR-ZP model on the C2 category, we observe a balanced accuracy of 0.73. Therefore, at least 46% of PPIs in the test set can be explained using PPI motifs learned from the C2 training set alone. This is a conservative estimate since we assume a perfect model that can predict PPI motifs from the training set with perfect accuracy. We chose to use the CoMET OH-SMR-ZP and not the superior NC-CoMET PC-AC since it enforces contiguous sequence motifs, thus we believe provides better intuition about their nature.

The results also show that there is only a small difference between models which enforce contiguous motifs and those who do not, which implies that at least the majority of PPI motifs are either contiguous or can be captured in a relatively short contiguous sequence. We also observed that the Auto-Covariance with physiochemical and SMR features representation showed superior performance throughout the analysis. The detector length analysis did not show significant differences in performance between models with detectors of various sizes, which suggests that this short contiguous sequence is not longer than 10-15 residues. Future work will investigate ways to improve the performance of the individual prediction models including deeper architectures, data augmentation based on homology, using PSSM features, and introducing attention mechanisms.

Chapter 5

Conclusion and Future Work

This thesis investigated the power of Deep Learning in the context of protein-protein interactions. A field which has suffered from sub-optimal standardization, a fact that prevented previous results from being translated into practice. We introduced a set of rigorous techniques for data curation, representation, and splitting, with the hope that these gold-standards will bridge the gap between computational results and their experimental and theoretical impact. Moreover, we developed NetPPI, a carefully curated dataset for development, training, and evaluation of PPI prediction models. Finally, we presented B-PPI, a novel deep learning approach to PPI predictions, which outperforms the previous state-of-the-art in a set prediction tasks.

Many opportunities arise from the practices and methods described in this work. The prediction models presented in this work can be significantly improved through experimentation with different machine learning techniques like attention networks, batch normalization, and data augmentation. This is especially important since a predictive model with very high accuracy is critical for many future applications. Such high-quality model can be used to perform sequence-based computational Alanine Scans, in which a known pair of interacting proteins will be evaluated multiple times, each time with a different sequence mutation. By analyzing the score differences between mutations, one might discover the most critical residues for the interactions (i.e., the hot spots). This approach presents vast advantages over experimental Alanine scans and structure-based computational scans. Computational scans are remarkably fast and cheap, thus allowing for many more mutations to be tested. It will also enable testing of multi-point mutations, which might reveal non-linear behaviors that are still unexplored.

A fascinating next step is a transition from PPI prediction to PPI design. Assuming a high-accuracy prediction model and by incorporating techniques like generative modeling and reinforcement learning, it should be possible to search and design an optimal PPI partner for a target of interest. Such capability has clear and immediate importance to the pharmaceutical industry, as well as other biological applications, as it offers a way to regulate PPI activities in living organisms through computationally-designed biological drugs. Finally, improved characterization of PPIs, down to the single residue resolution, might enable the development of small-molecule PPI regulator drugs.

Another exciting direction is the use of Deep Learning with structural information to characterize PPIs. At the moment, one of the main challenges is the lack of sufficient high-quality structures. A possible way to address this challenge is by thinking about interactions between protein domains as PPIs. It has been shown that in many cases, domains that were cleaved by proteases still come together and perform their respective function as if they were one protein. The reason for individual domains can act as a single protein lies in the same kind of interactions between residues that allow proteins to form complexes. Thus, we can take all known structures with more than one domain, and after "cutting" the inter-domain covalent bonds, treat them as PPIs. This approach will generate an order of magnitude more PPI structures and will support the development of structure-based Deep Learning methods. It will also help to investigate the hypothesis that PPIs originated in larger proteins that diverged through evolution into smaller building blocks.

Appendix A

Bayesian Hyperband

Performance of machine learning algorithms relies heavily on the configuration of hyperparameters such as learning rate, dropout probabilities, type of optimizer, and the loss function of choice. These hyperparameters form a complex, non-convex, multi-dimensional space and confront the designer of the model with a difficult task of finding an optimal configuration. The situation gets even more complicated for models that require long training times, as is the case for most deep neural networks. Hence, hyperparameter optimization is often referred to as the "black art" of machine learning. Conventional methods like grid and random [89] searches require excessive computation resources and domain-specific expertise, and produce sub-optimal results. In recent years, two novel approaches were suggested, Bayesian Optimization techniques [90] utilize results from historical configurations to inform sampling of new configurations, and Hyperband [91] improves results of random searches by allocating training resources to different configurations based on their performance. It was currently identified [92, 93] that the two methods are complementary and can be combined. In this appendix, we improve on Falkner et al. and Wang et al. by introducing an adaptation to their combined Bayesian Optimization and Hyperband model. We call this model Bayesian Hyperband (BHB). BHB was used for hyperparameter optimization of all deep-learning models in this thesis.

Algorithm 1 Wang et al. - Combination of Hyperband and Bayesian optimization

```
input: maximum resource budget that can be allocated to a single configuration $R$,
    and proportion constant $\eta$
output: optimal hyperparameter configuration
initialization: $s_{max} = \lfloor log_\eta(R) \rfloor$ , $B = (s_{max} + 1)R$
for $s \in \{s_{max}, s_{max} - 1, ..., 0\}$ do
    $n = \left\lceil \frac{B}{R} \frac{\eta^s}{(s+1)} \right\rceil$, $r = R\eta^{-s}$
    for $i \in 0, ..., s$ do
        $n_i = \lfloor n\eta^{-i} \rfloor$
        $r_i = r\eta^i$
        if $i == 0$ then
            $X = \emptyset$, $D_0 = \emptyset$
            for $t \in \{1, 2, ..., n_i\}$ do
                $x_{t+1} = argmax_x \mu(x|D_t)$
                $f(x_{t+1}) = \text{run\_then\_return\_obj\_val}(x, r_i)$
                $X = X \cup \{x_{t+1}\}$
                $D_{t+1} = D_t \cup \{(x_{t+1}, f(x_{t+1}))\}$
                Update probabilistic surrogate model using $D_{t+1}$
        else
            $F = \{ \text{ run\_then\_return\_obj\_val}(x, r_i) : x \in X\}$
            $X = \text{top\_k}(X, F, \lfloor n_i/\eta \rfloor)$
return optimal configuration
```

The method developed by Wang et al. is shown in Algorithm 1. Here the function $run\_then\_return\_obj\_val(x, r)$ evaluates the performance of a model with configuration x and a budget of r. The function $top\_k(trials, obj\_vals, k)$ is used for Hyperband's successive halving step. $\mu(x|D_t)$ models the expected improvement in the following equation:

$$\mu(x|D_t) = \mathbb{E}(max\{0, f_{t+1}(x) - f(x^+)\}|D_t) \tag{A.1}$$

, where x^+ is the optimal trial point after the first t steps. Lastly, $argmax_x \mu(x|D_t)$ is used to sample a new configuration with maximal expected improvement, based on the historical data. Following [94], a Tree-structured Parzen Estimator (TPE) is used to model $p(x|f(x))$ and $p(f(x))$.

Algorithm 2 Bayesian Hyperband

```
input: maximum resource budget that can be allocated to a single configuration $R$,
    proportion constant $\eta$, and constant fraction of random configurations $p$
output: optimal configuration
initialization: $s_{max} = \lfloor log_\eta(R) \rfloor$ , $B = (s_{max} + 1)R$, $D = \emptyset$
for $s \in \{0, 1, ..., s_{max}\}$ do
    $n = \left\lceil \frac{B}{R} \frac{\eta^s}{(s+1)} \right\rceil$, $r = R\eta^{-s}$
    for $i \in 0, ..., s$ do
        $n_i = \lfloor n\eta^{-i} \rfloor$
        $r_i = r\eta^i$
        if $i == 0$ then
            $X = \emptyset$
            for $t \in \{1, 2, ..., n_i\}$ do
                if rand() $< p$ then
                    $x_t$ = get_random_configuration()
                else
                    $x_t = argmax_x \mu(x|D)$
                $f(x_t)$ = run_then_return_obj_val($x, r_i$)
                $X = X \cup \{x_t\}$
                $D = D \cup \{(x_t, f(x_t))\}$
                Update probabilistic surrogate model using the updated $D$
        else
            $F = \{$ run_then_return_obj_val($x, r_i$) : $x \in X\}$
            $X$ = top_k($X, F, \lfloor n_i/\eta \rfloor$)
return optimal configuration
```

BHB (Algorithm 2) introduces several key improvements to the algorithm described by Wang et al. First, the constant p (first suggested by Falkner et al.) is used to uniformly sample random configurations at a given ratio. This provide the same theoretical guarantees of Hyperband, by ensuring that the algorithm is at most slower the Hyperband by a constant factor. Second, instead of resetting the trial dataset D for every new value of s, we keep a log of all trial configurations for future use, which allows us to constantly improve our surrogate model. Finally, we invert the order of training and start with the lowest budgets (i.e., values of s).

Appendix B

Methods

B.1 BLAST parameters

Parameter	Value
database	swissprot
E value	0.001
word size	6
gap open cost	11
gap extend cost	1
matrix	BLOSUM62
threshold	10
window size	40

Table B.1: BLAST-P parameters.

Parameter	Value
database	swissprot
E value	10
word size	5
gap open cost	11
gap extend cost	1
matrix	BLOSUM62
threshold	11
window size	40

Table B.2: PSI-BLAST parameters.

B.2 Representations

Amino acid	H_1	H_2	NCI	P_1	P_2	SASA	V
A	0.62	-0.5	0.007187	8.1	0.046	1.181	27.5
C	0.29	-1	-0.03661	5.5	0.128	1.461	44.6
D	-0.9	3	-0.02382	13	0.105	1.587	40
E	-0.74	3	0.006802	12.3	0.151	1.862	62
F	1.19	-2.5	0.037552	5.2	0.29	2.228	115.5
G	0.48	0	0.179052	9	0	0.881	0
H	-0.4	-0.5	-0.01069	10.4	0.23	2.025	79
I	1.38	-1.8	0.021631	5.2	0.186	1.81	93.5
K	-1.5	3	0.017708	11.3	0.219	2.258	100
L	1.06	-1.8	0.051672	4.9	0.186	1.931	93.5
M	0.64	-1.3	0.002683	5.7	0.221	2.034	94.1
N	-0.78	2	0.005392	11.6	0.134	1.655	58.7
P	0.12	0	0.239531	8	0.131	1.468	41.9
Q	-0.85	0.2	0.049211	10.5	0.18	1.932	80.7
R	-2.53	3	0.043587	10.5	0.291	2.56	105
S	-0.18	0.3	0.004627	9.2	0.062	1.298	29.3
T	-0.05	-0.4	0.003352	8.6	0.108	1.525	51.3
V	1.08	-1.5	0.057004	5.9	0.14	1.645	71.5
W	0.81	-3.4	0.037977	5.4	0.409	2.663	145.5
Y	0.26	-2.3	0.023599	6.2	0.298	2.368	117.3

Table B.3: Physiochemical properties of amino acids. H_1: hydrophobicity; H_2: hydrophilicity; NCI: net charge index of side chains; P_1: polarity; P_2: polarizability; SASA: solvent accessible surface area; V: volume of side chains; **Source: [37]**.

Cluster	Amino Acids
1	Ala, Gly, Val
2	Ile, Leu, Phe, Pro
3	Tyr, Met, Thr, Ser
4	His, Asn, Gln, Trp
5	Arg, Lys
6	Asp, Glu
7	Cys

Table B.4: Clustering of the 20 amino acids to seven clusters by their dipole and side chain volume. **Source: [40]**.

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	B	Z	X	*
A	4	-1	-2	-2	0	-1	-1	0	-2	-1	-1	-1	-1	-2	-1	1	0	-3	-2	0	-2	-1	0	-4
R	-1	5	0	-2	-3	1	0	-2	0	-3	-2	2	-1	-3	-2	-1	-1	-3	-2	-3	-1	0	-1	-4
N	-2	0	6	1	-3	0	0	0	1	-3	-3	0	-2	-3	-2	1	0	-4	-2	-3	3	0	-1	-4
D	-2	-2	1	6	-3	0	2	-1	-1	-3	-4	-1	-3	-3	-1	0	-1	-4	-3	-3	4	1	-1	-4
C	0	-3	-3	-3	9	-3	-4	-3	-3	-1	-1	-3	-1	-2	-3	-1	-1	-2	-2	-1	-3	-3	-2	-4
Q	-1	1	0	0	-3	5	2	-2	0	-3	-2	1	0	-3	-1	0	-1	-2	-1	-2	0	3	-1	-4
E	-1	0	0	2	-4	2	5	-2	0	-3	-3	1	-2	-3	-1	0	-1	-3	-2	-2	1	4	-1	-4
G	0	-2	0	-1	-3	-2	-2	6	-2	-4	-4	-2	-3	-3	-2	0	-2	-2	-3	-3	-1	-2	-1	-4
H	-2	0	1	-1	-3	0	0	-2	8	-3	-3	-1	-2	-1	-2	-1	-2	-2	2	-3	0	0	-1	-4
I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4	2	-3	1	0	-3	-2	-1	-3	-1	3	-3	-3	-1	-4
L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4	-2	2	0	-3	-2	-1	-2	-1	1	-4	-3	-1	-4
K	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5	-1	-3	-1	0	-1	-3	-2	-2	0	1	-1	-4
M	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5	0	-2	-1	-1	-1	-1	1	-3	-1	-1	-4
F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6	-4	-2	-2	1	3	-1	-3	-3	-1	-4
P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7	-1	-1	-4	-3	-2	-2	-1	-2	-4
S	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4	1	-3	-2	-2	0	0	0	-4
T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5	-2	-2	0	-1	-1	0	-4
W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11	2	-3	-4	-3	-2	-4
Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	-1	-3	-2	-1	-4
V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4	-3	-2	-1	-4
B	-2	-1	3	4	-3	0	1	-1	0	-3	-4	0	-3	-3	-2	0	-1	-4	-3	-3	4	1	-1	-4
Z	-1	0	0	1	-3	3	4	-2	0	-3	-3	1	-1	-3	-1	0	-1	-3	-2	-2	1	4	-1	-4
X	0	-1	-1	-1	-2	-1	-1	-1	-1	-1	-1	-1	-1	-1	-2	0	0	-2	-1	-1	-1	-1	-1	-4
*	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	1

Table B.5: BLOSUM62 Matrix.

Bibliography

[1] C A Finlay, P W Hinds, and A J Levine. The p53 proto-oncogene can act as a suppressor of transformation. *Cell*, 57(7):1083–93, jun 1989.

[2] D Malkin, F P Li, L C Strong, J F Fraumeni, C E Nelson, D H Kim, J Kassel, M A Gryka, F Z Bischoff, and M A Tainsky. Germ line p53 mutations in a familial syndrome of breast cancer, sarcomas, and other neoplasms. *Science (New York, N.Y.)*, 250(4985):1233–8, nov 1990.

[3] Shiv Srivastava, Zhiqiang Zou, Kathleen Pirollo, William Blattner, and Esther H. Chang. Germ-line transmission of a mutated p53 gene in a cancer-prone family with Li-Fraumeni syndrome. *Nature*, 348(6303):747–749, dec 1990.

[4] Ashutosh Malhotra, Erfan Younesi, Sudeep Sahadevan, Joerg Zimmermann, and Martin Hofmann-Apitius. Exploring novel mechanistic insights in Alzheimer's disease by assessing reliability of protein interactions. *Scientific Reports*, 5(1):13634, nov 2015.

[5] Glenn C Telling, Michael Scott, James Mastrianni, Ruth Gabizon, lI Marilyn Torchia, Fred E Cohen, Stephen J DeArmond, and Stanley B Prusiner. Prion Propagation in Mice Expressing Human and Chimeric PrP Transgenes Implicates the Interaction of Cellular PrP with Another Protein. *Cell*, 83:79–90, 1995.

[6] Heike Goehler, Maciej Lalowski, Ulrich Stelzl, Stephanie Waelter, Martin Stroedicke, Uwe Worm, Anja Droege, Katrin S Lindenberg, Maria Knoblich, Christian Haenig, et al. A protein interaction network links git1, an enhancer of huntingtin aggregation, to huntington's disease. *Molecular cell*, 15(6):853–865, 2004.

[7] Stanley Fields and Ok-kyu Song. A novel genetic system to detect protein–protein interactions. *Nature*, 340(6230):245–246, 1989.

[8] Valerie Benard and Gary M. Bokoch. Assay of Cdc42, Rac, and Rho GTPase Activation by Affinity Methods. *Methods in Enzymology*, 345:349–359, jan 2002.

[9] Ling Ren, Edith Chang, Khadijah Makky, Arthur L Haas, Barbara Kaboord, and M Walid Qoronfleh. Glutathione S-transferase pull-down assays using dehydrated immobilized glutathione resin. *Analytical Biochemistry*, 322(2):164–169, nov 2003.

[10] Frank Alber, Svetlana Dokudovskaya, Liesbeth M. Veenhoff, Wenzhu Zhang, Julia Kipper, Damien Devos, Adisetyantari Suprapto, Orit Karni-Schmidt, Rosemary Williams, Brian T. Chait, Michael P. Rout, and Andrej Sali. Determining the architectures of macromolecular assemblies. *Nature*, 450(7170):683–694, nov 2007.

[11] Bio-Resource: Yeast Two Hybrid System - For Protein - Protein Interaction Studies, http://technologyinscience.blogspot.com, 2014.

[12] James A. Wells. [18] systematic mutational analyses of protein-protein interfaces. In *Molecular Design and Modeling: Concepts and Applications Part A: Proteins, Peptides, and Enzymes*, volume 202 of *Methods in Enzymology*, pages 390 – 411. Academic Press, 1991.

[13] Jeffrey J Gray, Stewart Moughon, Chu Wang, Ora Schueler-Furman, Brian Kuhlman, Carol A Rohl, and David Baker. Protein-protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations. *Journal of molecular biology*, 331(1):281–99, aug 2003.

[14] Haidong Wang, Eran Segal, Asa Ben-Hur, Qian-Ru Li, Marc Vidal, and Daphne Koller. InSite: a computational method for identifying protein-protein interaction binding sites on a proteome-wide scale. 8(9), 2007.

[15] Chu Wang, Ora Schueler-Furman, and David Baker. Improved side-chain modeling for protein-protein docking. *Protein science : a publication of the Protein Society*, 14(5):1328–39, may 2005.

[16] Tanja Kortemme, David E Kim, and David Baker. Computational Alanine Scanning of Protein-Protein Interfaces. *Science Signaling*, 2004(219):pl2–pl2, feb 2004.

[17] R Manjunatha Kini and Herbert J Evans. Prediction of potential protein-protein interaction sites from amino acid sequence. *FEBS letters*, 385(1-2):81–86, 1996.

[18] J Willem M Nissink, Marcel L Verdonk, and Gerhard Klebe. Simple knowledge-based descriptors to predict protein-ligand interactions. methodology and validation. *Journal of computer-aided molecular design*, 14(8):787–803, 2000.

[19] Susan Jones and Janet M Thornton. Prediction of protein-protein interaction sites using patch analysis. *Journal of molecular biology*, 272(1):133–143, 1997.

[20] Florencio Pazos, Manuela Helmer-Citterich, Gabriele Ausiello, and Alfonso Valencia. Correlated mutations contain information about protein-protein interaction. *Journal of molecular biology*, 271(4):511–523, 1997.

[21] Edward M Marcotte, Matteo Pellegrini, Ho-Leung Ng, Danny W Rice, Todd O Yeates, and David Eisenberg. Detecting protein function and protein-protein interactions from genome sequences. *Science*, 285(5428):751–753, 1999.

[22] Martijn Huynen, Berend Snel, Warren Lathe, and Peer Bork. Predicting protein function by genomic context: quantitative evaluation and qualitative inferences. *Genome research*, 10(8):1204–1210, 2000.

[23] Florencio Pazos and Alfonso Valencia. Similarity of phylogenetic trees as indicator of protein–protein interaction. *Protein engineering*, 14(9):609–614, 2001.

[24] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015.

[25] Heng Wang, Zengchang Qin, and Tao Wan. Text generation based on generative adversarial nets with latent variable. *CoRR*, abs/1712.00170, 2017.

[26] Aäron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew W. Senior, and Koray Kavukcuoglu. Wavenet: A generative model for raw audio. *CoRR*, abs/1609.03499, 2016.

[27] M L Jul. Glow : Generative Flow. pages 1–15, jul 2018.

[28] Elizabeth Gibney. Google ai algorithm masters ancient game of go. *Nature News*, 529(7587):445, 2016.

[29] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin A. Riedmiller. Playing atari with deep reinforcement learning. *CoRR*, abs/1312.5602, 2013.

[30] John Markoff. Computer wins on 'jeopardy!': trivial, it's not. *New York Times*, 16, 2011.

[31] Babak Alipanahi, Andrew Delong, Matthew T Weirauch, and Brendan J Frey. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nature Biotechnology*, 33(8):831–838, jul 2015.

[32] Sheng Wang, Siqi Sun, Zhen Li, Renyu Zhang, and Jinbo Xu. Accurate de novo prediction of protein contact map by ultra-deep learning model. *PLoS computational biology*, 13(1):e1005324, 2017.

[33] Joel R. Bock and David A. Gough. Predicting protein–protein interactions from primary structure. *Bioinformatics (Oxford, England)*, 17(5):455–460, may 2001.

[34] Loris Nanni. Fusion of classifiers for predicting protein-protein interactions. *Neurocomputing*, 68(1-4):289–296, oct 2005.

[35] Loris Nanni. Hyperplanes for predicting protein-protein interactions. *Neurocomputing*, 69(1-3):257–263, dec 2005.

[36] J. Shen, J. Zhang, X. Luo, W. Zhu, K. Yu, K. Chen, Y. Li, and H. Jiang. Predicting protein-protein interactions based only on sequences information. *Proceedings of the National Academy of Sciences*, 104(11):4337–4341, mar 2007.

[37] Yanzhi Guo, Lezheng Yu, Zhining Wen, and Menglong Li. Using support vector machine combined with auto covariance to predict protein-protein interactions from protein sequences. *Nucleic Acids Research*, 36(9):3025–3030, may 2008.

[38] Yu-An Huang, Zhu-Hong You, Xin Gao, Leon Wong, and Lirong Wang. Using Weighted Sparse Representation Model Combined with Discrete Cosine Transformation to Predict Protein-Protein Interactions from Protein Sequence. *BioMed research international*, 2015:902198, oct 2015.

[39] Zhu-Hong You, Keith C. C. Chan, and Pengwei Hu. Predicting protein-protein interactions from primary protein sequences using a novel multi-scale local feature representation scheme and the random forest. *PloS one*, 10(5):e0125811, may 2015.

[40] Tanlin Sun, Bo Zhou, Luhua Lai, and Jianfeng Pei. Sequence-based prediction of protein protein interaction using a deep-learning algorithm. *BMC Bioinformatics*, 18(1):277, dec 2017.

[41] Yan-Bin Wang, Zhu-Hong You, Xiao Li, Tong-Hai Jiang, Xing Chen, Xi Zhou, and Lei Wang. Predicting protein-protein interactions from protein sequences by a stacked sparse autoencoder deep neural network. *Mol. BioSyst.*, 13(7):1336–1344, 2017.

[42] Asa Ben-Hur and William Stafford Noble. Choosing negative examples for the prediction of protein-protein interactions. *BMC bioinformatics*, 7(1):S2, 2006.

[43] Pawel Smialowski, Philipp Pagel, Philip Wong, Barbara Brauner, Irmtraud Dunger, Gisela Fobo, Goar Frishman, Corinna Montrone, Thomas Rattei, Dmitrij Frishman, et al. The negatome database: a reference set of non-interacting protein pairs. *Nucleic acids research*, 38(suppl_1):D540–D544, 2009.

[44] Daniel V Veres, Dávid M Gyurkó, Benedek Thaler, Kristóf Z Szalay, Dávid Fazekas, Tamás Korcsmáros, and Peter Csermely. Comppi: a cellular compartment-specific database for protein–protein interaction network analysis. *Nucleic acids research*, 43(D1):D485–D493, 2014.

[45] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 248–255. Ieee, 2009.

[46] Zhenqin Wu, Bharath Ramsundar, Evan N Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh S Pappu, Karl Leswing, and Vijay Pande. Moleculenet: a benchmark for molecular machine learning. *Chemical science*, 9(2):513–530, 2018.

[47] Gary D Bader, Michael P Cary, and Chris Sander. Pathguide: a pathway resource list. *Nucleic acids research*, 34(suppl_1):D504–D506, 2006.

[48] B Snel, G Lehmann, P Bork, and M A Huynen. STRING: a web-server to retrieve and display the repeatedly occurring neighbourhood of a gene. *Nucleic acids research*, 28(18):3442–4, sep 2000.

[49] Christian von Mering, Martijn Huynen, Daniel Jaeggi, Steffen Schmidt, Peer Bork, and Berend Snel. STRING: a database of predicted functional associations between proteins. *Nucleic acids research*, 31(1):258–61, jan 2003.

[50] C. von Mering, Lars J Jensen, Berend Snel, Sean D Hooper, Markus Krupp, Mathilde Foglierini, Nelly Jouffre, Martijn A Huynen, and Peer Bork. STRING: known and predicted protein-protein associations, integrated and transferred across organisms. *Nucleic Acids Research*, 33(Database issue):D433–D437, dec 2004.

[51] C. von Mering, L. J. Jensen, M. Kuhn, S. Chaffron, T. Doerks, B. Kruger, B. Snel, and P. Bork. STRING 7–recent developments in the integration and prediction of protein interactions. *Nucleic Acids Research*, 35(Database):D358–D362, jan 2007.

[52] L. J. Jensen, M. Kuhn, M. Stark, S. Chaffron, C. Creevey, J. Muller, T. Doerks, P. Julien, A. Roth, M. Simonovic, P. Bork, and C. von Mering. STRING 8–a global view on proteins and their functional interactions in 630 organisms. *Nucleic Acids Research*, 37(Database):D412–D416, jan 2009.

[53] D. Szklarczyk, A. Franceschini, M. Kuhn, M. Simonovic, A. Roth, P. Minguez, T. Doerks, M. Stark, J. Muller, P. Bork, L. J. Jensen, and C. v. Mering. The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic Acids Research*, 39(Database):D561–D568, jan 2011.

[54] Andrea Franceschini, Damian Szklarczyk, Sune Frankild, Michael Kuhn, Milan Simonovic, Alexander Roth, Jianyi Lin, Pablo Minguez, Peer Bork, Christian von Mering, and Lars J. Jensen. STRING v9.1: protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Research*, 41(D1):D808–D815, nov 2012.

[55] Damian Szklarczyk, Andrea Franceschini, Stefan Wyder, Kristoffer Forslund, Davide Heller, Jaime Huerta-Cepas, Milan Simonovic, Alexander Roth, Alberto Santos, Kalliopi P. Tsafou, Michael Kuhn, Peer Bork, Lars J. Jensen, and Christian von Mering. STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Research*, 43(D1):D447–D452, jan 2015.

[56] Andrea Franceschini, Jianyi Lin, Christian von Mering, and Lars Juhl Jensen. SVD-phy: improved prediction of protein functional associations through singular value decomposition of phylogenetic profiles. *Bioinformatics*, 32(7):1085–1087, apr 2016.

[57] Damian Szklarczyk, John H Morris, Helen Cook, Michael Kuhn, Stefan Wyder, Milan Simonovic, Alberto Santos, Nadezhda T Doncheva, Alexander Roth, Peer Bork, Lars J. Jensen, and Christian von Mering. The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible. *Nucleic Acids Research*, 45(D1):D362–D368, jan 2017.

[58] K. R. Brown and I. Jurisica. Online Predicted Human Interaction Database. *Bioinformatics*, 21(9):2076–2082, may 2005.

[59] Kevin R Brown and Igor Jurisica. Unequal evolutionary conservation of human protein interactions in interologous networks. *Genome Biology*, 8(5):R95, 2007.

[60] C. Stark, Bobby-Joe Breitkreutz, Teresa Reguly, Lorrie Boucher, Ashton Breitkreutz, and Mike Tyers. BioGRID: a general repository for interaction datasets. *Nucleic Acids Research*, 34(90001):D535–D539, jan 2006.

[61] Sandra Orchard, Mais Ammari, Bruno Aranda, Lionel Breuza, Leonardo Briganti, Fiona Broackes-Carter, Nancy H Campbell, Gayatri Chavali, Carol Chen, Noemi Del-Toro, Margaret Duesbury, Marine Dumousseau, Eugenia Galeota, Ursula Hinz, Marta Iannuccelli, Sruthi Jagannathan, Rafael Jimenez, Jyoti Khadake, Astrid Lagreid, Luana Licata, Ruth C Lovering, Birgit Meldal, Anna N Melidoni, Mila Milagros, Daniele Peluso, Livia Perfetto, Pablo Porras, Arathi Raghunath, Sylvie Ricard-Blum, Bernd Roechert, Andre Stutz, Michael Tognolli, Kim van Roey, Gianni Cesareni, and Henning Hermjakob. The MIntAct project–IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic acids research*, 42(Database issue):D358–63, jan 2014.

[62] Diego Alonso-López, Miguel A. Gutiérrez, Katia P. Lopes, Carlos Prieto, Rodrigo Santamaría, and Javier DeLasRivas. APID interactomes: providing proteome-based interactomes with controlled quality for multiple species and derived networks. *Nucleic Acids Research*, 44(W1):W529–W535, jul 2016.

[63] Taibo Li, Rasmus Wernersson, Rasmus B Hansen, Heiko Horn, Johnathan Mercer, Greg Slodkowicz, Christopher T Workman, Olga Rigina, Kristoffer Rapacki, Hans H Stærfeldt, Søren Brunak, Thomas S Jensen, and Kasper Lage. A scored human protein-protein interaction network to catalyze genomic interpretation. *Nature Methods*, 14(1):61–64, jan 2017.

[64] Martin H. Schaefer, Jean-Fred Fontaine, Arunachalam Vinayagam, Pablo Porras, Erich E. Wanker, and Miguel A. Andrade-Navarro. HIPPIE: Integrating Protein Interaction Networks with Experiment Based Quality Scores. *PLoS ONE*, 7(2):e31826, feb 2012.

[65] Martin H. Schaefer, Tiago J. S. Lopes, Nancy Mah, Jason E. Shoemaker, Yukiko Matsuoka, Jean-Fred Fontaine, Caroline Louis-Jeune, Amie J. Eisfeld, Gabriele Neumann, Carol Perez-Iratxeta, Yoshihiro Kawaoka, Hiroaki Kitano,

and Miguel A. Andrade-Navarro. Adding Protein Context to the Human Protein-Protein Interaction Network to Reveal Meaningful Interactions. *PLoS Computational Biology*, 9(1):e1002860, jan 2013.

[66] Apichat Suratanee, Martin H. Schaefer, Matthew J. Betts, Zita Soons, Heiko Mannsperger, Nathalie Harder, Marcus Oswald, Markus Gipp, Ellen Ramminger, Guillermo Marcus, Reinhard Männer, Karl Rohr, Erich Wanker, Robert B. Russell, Miguel A. Andrade-Navarro, Roland Eils, and Rainer König. Characterizing Protein Interactions Employing a Genome-Wide siRNA Cellular Phenotyping Screen. *PLoS Computational Biology*, 10(9):e1003814, sep 2014.

[67] Gregorio Alanis-Lobato, Miguel A. Andrade-Navarro, and Martin H. Schaefer. HIPPIE v2.0: enhancing meaningfulness and reliability of protein-protein interaction networks. *Nucleic Acids Research*, 45(D1):D408–D414, jan 2017.

[68] Luana Licata, Leonardo Briganti, Daniele Peluso, Livia Perfetto, Marta Iannuccelli, Eugenia Galeota, Francesca Sacco, Anita Palma, Aurelio Pio Nardozza, Elena Santonico, et al. Mint, the molecular interaction database: 2012 update. *Nucleic acids research*, 40(D1):D857–D861, 2011.

[69] L. Salwinski, Christopher S Miller, Adam J Smith, Frank K Pettit, James U Bowie, and David Eisenberg. The Database of Interacting Proteins: 2004 update. *Nucleic Acids Research*, 32(90001):449D–451, jan 2004.

[70] I Xenarios, D W Rice, L Salwinski, M K Baron, E M Marcotte, and D Eisenberg. DIP: the database of interacting proteins. *Nucleic acids research*, 28(1):289–91, jan 2000.

[71] I Xenarios, E Fernandez, L Salwinski, X J Duan, M J Thompson, E M Marcotte, and D Eisenberg. DIP: The Database of Interacting Proteins: 2001 update. *Nucleic acids research*, 29(1):239–41, jan 2001.

[72] Ioannis Xenarios, Lukasz Salwínski, Xiaoqun Joyce Duan, Patrick Higney, Sul-Min Kim, and David Eisenberg. DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. *Nucleic acids research*, 30(1):303–5, jan 2002.

[73] T. S. Keshava Prasad, R. Goel, K. Kandasamy, S. Keerthikumar, S. Kumar, S. Mathivanan, D. Telikicherla, R. Raju, B. Shafreen, A. Venugopal, L. Balakrishnan, A. Marimuthu, S. Banerjee, D. S. Somanathan, A. Sebastian, S. Rani, S. Ray, C. J. Harrys Kishore, S. Kanth, M. Ahmed, M. K. Kashyap, R. Mohmood, Y. L. Ramachandra, V. Krishna, B. A. Rahiman, S. Mohan, P. Ranganathan, S. Ramabadran, R. Chaerkady, and A. Pandey. Human Protein Reference Database–2009 update. *Nucleic Acids Research*, 37(Database):D767–D772, jan 2009.

[74] G. R. Mishra, M Suresh, K Kumaran, N Kannabiran, Shubha Suresh, P Bala, K Shivakumar, N Anuradha, Raghunath Reddy, T Madhan Raghavan, Shalini

Menon, G Hanumanthu, Malvika Gupta, Sapna Upendran, Shweta Gupta, M Mahesh, Bincy Jacob, Pinky Mathew, Pritam Chatterjee, K S Arun, Salil Sharma, K N Chandrika, Nandan Deshpande, Kshitish Palvankar, R Raghavnath, R Krishnakanth, Hiren Karathia, B Rekha, Rashmi Nayak, G Vishnupriya, H G Mohan Kumar, M Nagini, G S Sameer Kumar, Rojan Jose, P Deepthi, S Sujatha Mohan, T K B Gandhi, H C Harsha, Krishna S Deshpande, Malabika Sarker, T S Keshava Prasad, and Akhilesh Pandey. Human protein reference database–2006 update. *Nucleic Acids Research*, 34(90001):D411–D414, jan 2006.

[75] S. Peri, J Daniel Navarro, Ramars Amanchy, Troels Z Kristiansen, Chandra Kiran Jonnalagadda, Vineeth Surendranath, Vidya Niranjan, Babylakshmi Muthusamy, T K B Gandhi, Mads Gronborg, Nieves Ibarrola, Nandan Deshpande, K Shanker, H N Shivashankar, B P Rashmi, M A Ramya, Zhixing Zhao, K N Chandrika, N Padma, H C Harsha, A J Yatish, M P Kavitha, Minal Menezes, Dipanwita Roy Choudhury, Shubha Suresh, Neelanjana Ghosh, R Saravana, Sreenath Chandran, Subhalakshmi Krishna, Mary Joy, Sanjeev K Anand, V Madavan, Ansamma Joseph, Guang W Wong, William P Schiemann, Stefan N Constantinescu, Lily Huang, Roya Khosravi-Far, Hanno Steen, Muneesh Tewari, Saghi Ghaffari, Gerard C Blobe, Chi V Dang, Joe G N Garcia, Jonathan Pevsner, Ole N Jensen, Peter Roepstorff, Krishna S Deshpande, Arul M Chinnaiyan, Ada Hamosh, Aravinda Chakravarti, and Akhilesh Pandey. Development of Human Protein Reference Database as an Initial Platform for Approaching Systems Biology in Humans. *Genome Research*, 13(10):2363–2371, oct 2003.

[76] Andrei L Turinsky, Sabry Razick, Brian Turner, Ian M Donaldson, and Shoshana J Wodak. Literature curation of protein interactions: measuring agreement across major public databases. *Database*, 2010, 2010.

[77] Leonardo G Trabuco, Matthew J Betts, and Robert B Russell. Negative protein-protein interaction datasets derived from large-scale two-hybrid experiments. *Methods*, 58(4):343–348, 2012.

[78] Guillaume Launay, Nicoletta Ceres, and Juliette Martin. Non-interacting proteins may resemble interacting proteins: prevalence and implications. *Scientific Reports*, 7(1):40419, dec 2017.

[79] Stephen F. Altschul, Warren Gish, Webb Miller, Eugene W. Myers, and David J. Lipman. Basic local alignment search tool. *Journal of Molecular Biology*, 215(3):403–410, oct 1990.

[80] The UniProt Consortium. UniProt: the universal protein knowledgebase. *Nucleic Acids Research*, 45(D1):D158–D169, 2017.

[81] Peter J A Cock, Tiago Antao, Jeffrey T Chang, Brad A Chapman, Cymon J Cox, Andrew Dalke, Iddo Friedberg, Thomas Hamelryck, Frank Kauff, Bartek Wilczynski, and Michiel J L de Hoon. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, 25(11):1422–1423, 2009.

[82] Yaou Zhao, Yuehui Chen, and Mingyan Jiang. Predicting protein-protein interactions from protein sequences using probabilistic neural network and feature combination. *Journal of Information & Computational Science*, 11(7):2397–2406, 2014.

[83] Zhu-Hong You, Ying-Ke Lei, Lin Zhu, Junfeng Xia, and Bing Wang. Prediction of protein-protein interactions from amino acid sequences with ensemble extreme learning machines and principal component analysis. In *BMC bioinformatics*, volume 14, page S10. BioMed Central, 2013.

[84] L Breiman. Bagging predictors machine learning, 24, 123–140. 1996.

[85] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[86] Tsung-Yu Lin. Bilinear CNN Models for Fine-grained Visual Recognition.

[87] Thrasyvoulos Karydis. Learning hierarchical motif embeddings for protein engineering, Massachusetts Institute of Technology, 2017.

[88] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely Connected Convolutional Networks. aug 2016.

[89] James Bergstra and Yoshua Bengio. Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, 13(Feb):281–305, 2012.

[90] Jasper Snoek, Hugo Larochelle, and Ryan P Adams. Practical bayesian optimization of machine learning algorithms. In *Advances in neural information processing systems*, pages 2951–2959, 2012.

[91] Lisha Li, Kevin Jamieson, Giulia DeSalvo, Afshin Rostamizadeh, and Ameet Talwalkar. Hyperband: A novel bandit-based approach to hyperparameter optimization. *arXiv preprint arXiv:1603.06560*, 2016.

[92] Stefan Falkner, Aaron Klein, and Frank Hutter. Combining hyperband and bayesian optimization. In *Advances in neural information processing systems*, 2017.

[93] Jiazhuo Wang, Jason Xu, and Xuejun Wang. Combination of hyperband and bayesian optimization for hyperparameter optimization in deep learning. *CoRR*, abs/1801.01596, 2018.

[94] James S Bergstra, Rémi Bardenet, Yoshua Bengio, and Balázs Kégl. Algorithms for hyper-parameter optimization. In *Advances in neural information processing systems*, pages 2546–2554, 2011.