Faculty of Science and Technology
of the University of Coimbra

Strategies of Autoencoders in the Prediction of Protein-Protein Interactions

João Miguel Loureiro Albuquerque Antunes

Master Thesis in the field of Bioinformatics and Clinical Informatics
of Biomedical Engineering, mentored by Joel Perdiz Arrais and Carlos Pereira
in the Faculty of Science and Technology of the University of Coimbra

September 2019

Agradecimentos

Em primeiro lugar, não posso deixar de agradecer aos meus orientadores, Professor Carlos Pereira e Professor Joel Arrais, pelo acompanhamento que me deram, e pela exigência e profissionalismo que sempre tiveram comigo e me pediram em retorno. Em especial, ao Professor Joel, pela oportunidade de trabalhar neste projeto.

Aos meus colegas e amigos de curso, que desde o primeiro dia são a melhor coisa que este me pode alguma vez dar, e nos quais encontrei uma segunda família e até companheiros de casa, especialmente à Ana, ao Francisco, à Mafalda, à Mariana, ao Nelson, ao Oshley, e ao Malafaia.

Aos companheiros de viagem no associativismo, em especial ao Daniel, por acreditarem e fazerem de mim muito mais do que aquilo que sonhava ser, e me ajudaram a ajudar os outros, em especial os estudantes que com tanto orgulho servimos. À Maria por toda a ajuda num percurso tão desafiante do NEDF. À Ana Rita, ao André e ao Bento, por nos tornarmos Mestres em debater toda a Academia, enquanto nos vamos tornando Mestres nos respetivos cursos.

À Filipa e à Fradique, por serem as amigas que sendo um ano mais velhas têm o dom especial de me dar lições contantes sobre a vida, que tantas mais vezes deveria ouvir.
À Mafalda e à Rita que espero nunca ouvirem as minhas lições de vida, mas que sabem que as acompanharei sempre que precisarem.
À Inês, pela capacidade de equilíbrio entre a amizade, preocupação, paciência e tentativa constante de me incutir um juízo que nunca ganhei.

A um sarcasmo de crítica tão acertado e humorista do Zé e ao sempre responsável mano velho Guilherme, por saber que com eles poderei contar toda a vida.
Aos meus pais, confiantes supremos das minhas capacidades, tentando sempre manter os meus pés na terra, e perdoando com mais ou menos facilidade as faltas que para com eles tenho e que nunca deles tive, por tudo quanto sou e faço que se deve tanto a eles.

A ti Coimbra, por me acolheres durante estes cinco anos, e seres a minha casa num futuro tão incerto do qual espero nada menos do que te devolver tudo o que me deste, como a esta casa que é a Universidade de Coimbra.

Em especial à Associação Académica de Coimbra, eterno símbolo da liberdade e expressão estudantil, que me surpreende todos os dias que tenho a honra de ser sua pequena parte. Não trocava nada pelo que vivi contigo, e não seria metade do que sou se não te tivesse conhecido no NEDF, nas ruas, na Capa e Batina, e em tudo o que me envolveste nestes anos.

Resumo

A compreensão do corpo humano tem sido uma luta desde que a humanidade desenvolveu um sentido crítico à vida. As proteínas são um elemento chave no normal funcionamento do nosso organismo e intervém na maioria dos processos biológicos que existem em todos os seres vivos. A sua associação, designada Interação Proteína-Proteína, pode regular a sua expressão, mas também aumentar o seu número de funções, bem como maximizar ou reduzir o seu impacto.

Estas interações são de incalculável valor quando considerado o fator benéfico que podem ter para a crua de doenças e prevenção das mesmas.

No contexto da Descoberta e Implementação de Fármacos, o custo e tempo de síntese de uma nova molécula viável, que sobreviva aos ensaios clínicos e à aprovação das agências reguladoras, afirma-se como um dos problemas centrais para a indústria farmacêutica que se tem voltado para a Bioinformática como uma alternativa mais rápida e barata.

A pesquisa realizada nesta tese pretende ser a base de contexto no qual os Alvos Terapêuticos se devem desenvolver para alcançar esta redução de custos. A abordagem aplica Algoritmos de *Deep Learning*, especificamente *Autoencoders* e Máquinas de Vetores de Suporte, a conjuntos de dados de interações conhecidas e é capaz de reconhecer novos pares de proteínas que interagem num mesmo organismo ou no contexto interespécies.

Apesar dos métodos não serem novos, esta abordagem introduz o conceito de uso de erros no *encoding* dos *Autoencoders* como dimensões para classificação das entradas como pares de proteínas que interagem ou não interagem.

Os resultados preliminares mostram um elevado AUC na previsão, alcançando 0.970 para o organismo humano, mas com alguns desvios não justificados quando considerada a relação filogenética que precisam de ser analisados em trabalho futuro dentro da mesma espécie e em contexto interespécie.

Palavras-chave: Bioinformática, Desenvolvimento de Fármacos, Interação Proteína-Proteína, *Autoencoders*

Abstract

The understanding of the human body has always been a struggle since humanity has developed a critical sense to life. Proteins are a key element in the normal functioning of our organism and intervene in most of the biological processes that exist in all living beings. Their association, designated Protein-Protein Interactions, can regulate their expression but also increase the number of functions, as well as maximizing or reducing their impact.

These interactions are of invaluable worth when considering the benefic factor they can have in the healing of diseases and their prevention.

In the context of Drug Discovery and Deployment, the cost and time of synthesizing a new viable molecule, which survives clinical trials and regulatory agencies approval, stand as the core problems for the pharmaceutical industry which has turned the focus to Bioinformatics as a possible cost and time saving alternative.

The research made in this thesis intends to be the base of context to which the Drug Targeting must be developed to reach this goal of cost reduction. The approach applies Deep Learning Algorithms, specifically Autoencoders and Support Vector Machines, to datasets of known interactions and is able to discover new protein pairs that interact within the same organism or interspecies.

While the methods are not new, this approach introduces a concept of the use of the errors in the prediction of the Autoencoders as features to classify the inputs as Interacting or Non-Interacting pairs of proteins.

The preliminary results show high AUC in the prediction reaching 0.970 for the human organism, but with some unjustified deviations considering the phylogenetic relationships that need to be analyzed in future work both within the same species and interspecies trials.

Key-words: Bioinformatics, Drug Development, Protein-Protein Interactions, Autoencoders

Acknowledgments

This research has been funded by the Portuguese Research Agency FCT, through D4 - Deep Drug Discovery and Deployment (CENTRO-01-0145-FEDER-029266).

Contents

Index of figures

Abreviations

AC	Autocovariance
AE	Autoencoder
AUC	Area Under the Curve
CT	Conjoint Triad
CTE	Classifier Training Experiment
DNA	Deoxyribonucleic Acid
DRE	Dimension Reduction Experiment
ELM	Extreme Learning Machines
EME	Encoding Method Experiment
FN	False Negative
FP	False Positive
HL	Hidden layer
ID	Identifier
K-NN	K Nearest Neighbors
LDA	Linear Discriminant Analysis
LogSig	Logistic Sigmoid
LSTM	Long Short Term Memory
PPI	Protein-Protein Interactions
RDF	Random Forest
RNA	Ribonucleic Acid
ROC	Receiver Operating Characteristic
SAE	Stacked Autoencoder
SatLin	Saturation Linear
SVM	Suppot Vector Machines
TN	True Negative
TP	True Positive

1 Introduction

Humanity has always focused efforts to the development of therapies that can endure its resistance to harmful chemical agents present in nature and how its consumption could increase the quality of life and its durability.
The evolution of botany made the study focused on the biological characteristics of each organism, and the later development of bioinformatics has revolutionized what is known as the Drug Development Process [1].

Drug Development is the set of activities in which natural or artificial molecules are developed, tested and approved as therapeutic agents [2]. This process has turned from the biological search to the artificial synthesis considering the evolution of the Interaction Models, which can reduce the number of molecules to be tested in the earlier stages, as well as the amount of raw materials necessary for the experiments.

Proteins are the backbone for most of the biological processes that are developed in any living organisms, as well as most of the structural functions. Without their existence, and the interactions they establish, life would not be possible. These keen macromolecules, composed of amino acid sequences, are therefore the target of many investigators in the field of chemistry, biochemistry, biology, medicine and bioinformatics.
Proteins interact with DNA using their ability to replicate or inhibit the expression of a certain gene, with carbohydrates and lipids conditioning the way they are processed, with molecules such as oxygen allowing cells to perform metabolic pathways, and with other proteins creating complexes that perform millions of different functions [3].

Bioinformatics, being the field responsible for the connection of raw biological data to specific knowledge about a given biological process, is also focused in a substantial share of its efforts to proteins and their interactions, intending to reduce the time and resources needed for the detection of key elements in a protein interaction and the function that each protein and the interaction between them perform in a given environment.
Through gene and protein sequencing, expression and detection of interactions, Bioinformatics has been able to reduce the time and cost of the study of several processes by simulating based on already known processes and increasing the confidence of positive predictions as likely to be confirmed in laboratory procedures [4].
However, understanding the results and models that Bioinformatics gives is still a barrier to the development of other areas, or the full application of the findings these experiments produce.

Protein Interactions are one of the main topics of research by Bioinformatics, concerning the relationship between proteins and the way those interactions generate repercussions in the biochemical pathways in the human body [5].
These interactions are as important as proteins themselves, amplifying, diminishing or creating new functions through their actions. Protein interactions are what allows for the indirect interference of synthesized molecules in the human body performing the three types of effects mentioned.

These general concepts are the basis for the work developed in this thesis. From them, all the hypothesis and concepts in this document were developed to ensure that the final focus of drug development optimization is kept.

The first concept recreated, also essential to the understanding of the work developed, is the Pipeline. Although it is used primarily in Drug Development, it reaches areas which do not interfere with the specific subject of this theme. The Pipeline is the flow model created to ensure that every trained network and every chosen parameter in the set of experiments created is not lost. As components part of the Pipeline are, by order, the dataset, Encoding Method, Dimension Reduction Architecture, Classifier and Metrics of evaluation. The variation of these models is the general study object in which all experiments are inserted.

As final result of the work developed, a family of models for the Pipeline was created, resorting to computational methods that evaluate the probability of Protein Pairs binding in different contexts of the ones already found, or in the same contexts but with different connections. The identified interactions using this Pipeline are then applied to Drug Development, with the intention of maximizing accuracy and efficiency in the discovery of new molecules with therapeutic applications.

1.1 Motivation

The use of Bioinformatics and Computational Models in the context of protein interactions is still a recent field but with stable growth in search, sponsored by Research Centers and Drug Research Companies. The D4 project is a European Project focused on Deep Drug Discovery and Deployment to which Computational Models are the main focus [6].

Nowadays, Drug Development is mainly focused on new molecules synthesized through laboratory processes opposing to the classical drug development techniques that consisted in applying new found organism properties, especially outside of the *Animalia* kingdom. New plants, fungus and

bacteria provided for most of the medicine until the massification of the laboratory process now implemented, a cheaper method compared to the necessary endeavors to find new natural solution, result of the increased level of knowledge about drugs and species.

This process is however still expensive, with Drug Development cost statistics showing that, for each good, viable molecule, as much as 16,000 molecules and 2.87 billion dollars are spent [7]. Besides the economic aspect of this problem, we need still to consider the scarcity of some of the raw material to build these molecules, which make lab tests impossible in large quantities for the cases that the elements necessary are not abundant on Earth or need a substantial level of reshaping to become appealing to Drug Development [8].

D4 project intends to simplify Drug Development in the early stages of molecule testing, creating computational models that successfully reduce the number of molecules to be tested in the process. In a single hand, both biologic and economic resources are spared to intensive and deregulated use, when a model predicts that certain pair of proteins will surely not comply with each other.
This work intends to understand the relationship between macromolecules, proteins, that can feed important information to protein-drug interaction models later in the Drug Development Pipeline.

1.2 Objectives

The main goal is to develop a computational model for protein-protein interactions, giving the necessary flexibility and know how to also allow the construction of other models, by creating a conceptual line of model design for interaction prediction. To reach this goal, several steps were followed:

1. Study of Machine Learning approaches to protein-protein interaction prediction;
2. Study of biological processes, attaining the foundational properties of proteins and amino acids to be selected as features;
3. Dataset and feature selection, and encoding strategies applied to protein and amino acid properties;
4. Study of Autoencoders structures and its different variations;
5. Model training and testing in benchmark datasets;
6. Analysis of the model in interspecies context.

These steps are described in the chapters of this thesis intending to show the workflow of this thesis in the past year of work.
The final goal is to achieve a family of models that can predict protein interactions, considering the organism it relates to, and allow for the information to be used in benchmark contributions in interspecies interactions, which can open the way for new protein interaction networks, contribute for drug development as new test raw materials, and using already known molecules as resource for other possible interactions.

1.3 Document Structure

To improve the understanding of the document throughout its reading, this chapter presents this Master Thesis structure, a description to each Chapter and the phase of work it complies with.

Chapter 1. Introduction – This chapter contextualizes the Thesis in the field of Bioinformatics and Protein Interaction, and the professional characteristics.

Chapter 2. 1 Protein-Protein Interaction Problem – Contextualizes the Protein-Protein Interaction problem in a biological background approach, as a support for the knowledge attained to the computational phase.

Chapter 3. State of the Art for Protein interaction Detections – An overview of the previous and proposed methods, with the description of its structure, and the benefits or prejudices regarding the problem of protein interactions.

Chapter 4. datasets and Methods – Description of the initial datasets and preprocessing methods, as well as the used algorithms and specific precautions taken in the context of the problem.

Chapter 5. Experiments and Results – This chapter focus on the workflow in each experiment performed, as well as the outcomes they create to the next experiments.

Chapter 6. Discussion and Conclusions – Interpretation of the results obtained, and the data collected in the experiments as a mean to achieve the intended purpose of this thesis.

Chapter 7. Final Considerations and Future Work – Suggestions of improvement on the work developed in the thesis, as well as presentation of the options to continue this work in the same or other directions.

2 Protein-Protein Interaction Problem

The problem described in this chapter, the one that this work intends to reflect upon, is a classification problem. We intend to receive, through a designed model, the result of the following function:

$$y = interactionPredictor(features(p1), features(p2)) \qquad (1)$$

In this scenario, *interactionPredictor* is the model and the features to select will depend on the study based on biological and computational backgrounds. Proteins are represented by p1 and p2, arbitrarily. In the design of the model and experiments, the Drug Development was approached and studied as a base.

2.1 Drug Development

Drug Development, illustrated by Figure 1, is the general process of drug discovery and approval, which can lead up to 10.5 years and require investments of minimum 1.8 billion dollars [9]. These facts are of substantial worrisome, since the insurgence of new diseases, especially the ones that are transmittable, have enormous impacts in social and economic aspects of society, augmented in developing countries.

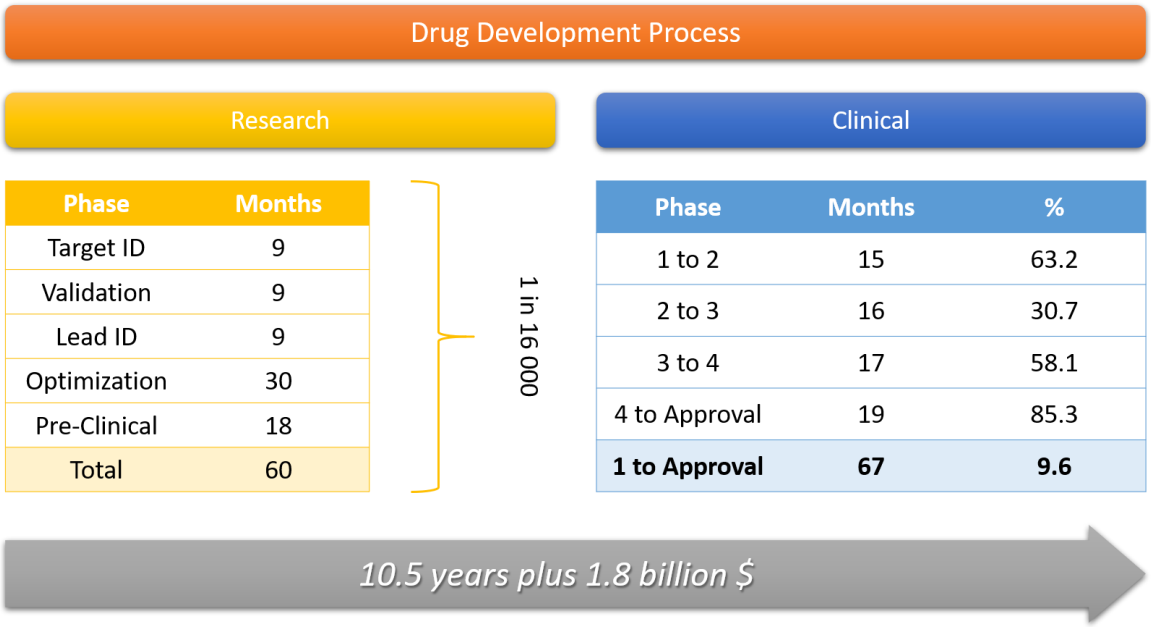

Figure 1 Drug Development General Process.

In Figure 1, the Research and Clinical Stage are divided, since the area of actuation is different from development to regulatory stage, respectively. Target ID is the process in which the researcher manages to locate the molecule that is related to the disease intended to study, validated in the second step. In the following stage, a Lead is identified to reach to that protein, suffering a process of optimization until optimum connection between Lead and Target.
The Clinical Stage focus on the approval by regulatory agencies, with Phase I being an adaption of the developed drug to the adequate doses in healthy individuals, Phase II as the first trials in unhealthy individuals, with small numbers of volunteers. This stage, as it can be noticed, is the one with the lowest approval rating. The following Phase III is a conceptual validation in larger number of patients, and Phase IV can be introduced in already approved drugs to monitor the evolution of the patients, in a market scale.

The Drug Development Research Stage, schematized in Figure 2, is the focus of the present thesis, in which the problem described by Equation 1 will have an impact. With a 5 year mean duration, it is the focused Stage by Bioinformatics and Pharmaceutical Industries, since Drug Trials are required and of low flexibility.
With a success of 1 in 16,000 tested molecules to be viable for Drug Trials, Research Stage has been continuously improved by the assistance of Computational Methods. The sequencing of the human genome, for example, enable Biomarkers, one of the strategies that increased the success rate of Drug Development, with a difference from non-Biomarker to Biomarker use of 8.4% to 25.9%, respectively [9].

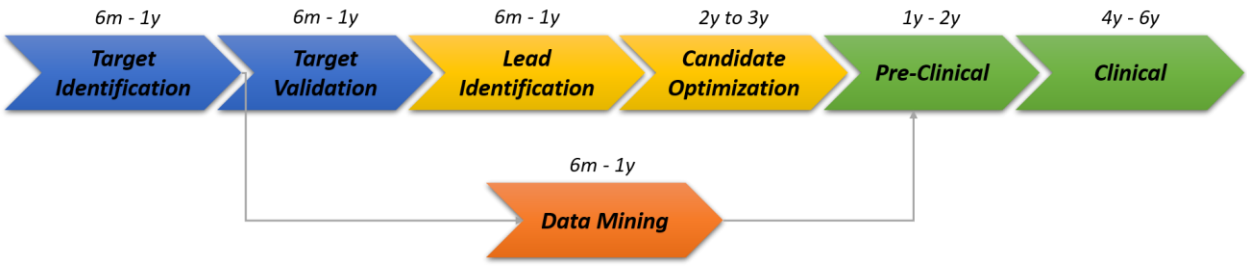

Figure 2 Drug Discovery Overview with and without Data Mining assistance, representing its main 6 steps, that are grouped in the two main stages previously mentioned

In this Stage, we can differentiate several types of tested molecules, which are then submitted to Drug Trials:

- New Molecular Entities – new molecules synthesized
- Biologic Entities – derived from Nature or naturally produced substances
- Non-New Molecular Entities – adaptation of drugs to new disease

These different types of entities have different approval rates, ranging from 6.2% to 11.5% and 22.6%, respectively [9]. Is in the Biological Entities that Protein-Protein Interactions have the

biggest and more direct impact, since they have direct influence in all biological environments, and their role in the normal functioning of every organism. However, the models created in Protein-Protein Interactions are of positive influence in models of Drug Target Interaction.

2.2 Proteins

Proteins are macromolecules composed by chains of amino acid sequences that perform most of the biological processes in any organism. They are created through processes of transcription and translation.

As the protein evolves and adds amino acids to its sequence, a more pressuring environment is created surrounding the said sequence, pushing it to fold and conform according to the micro biochemical conditions that each environment and amino acid connected creates [10].

These conformations, folding and otherwise movements and primary interactions between amino acids create different levels of structure, from primary to quaternary protein structure, as shown in Figure 3.

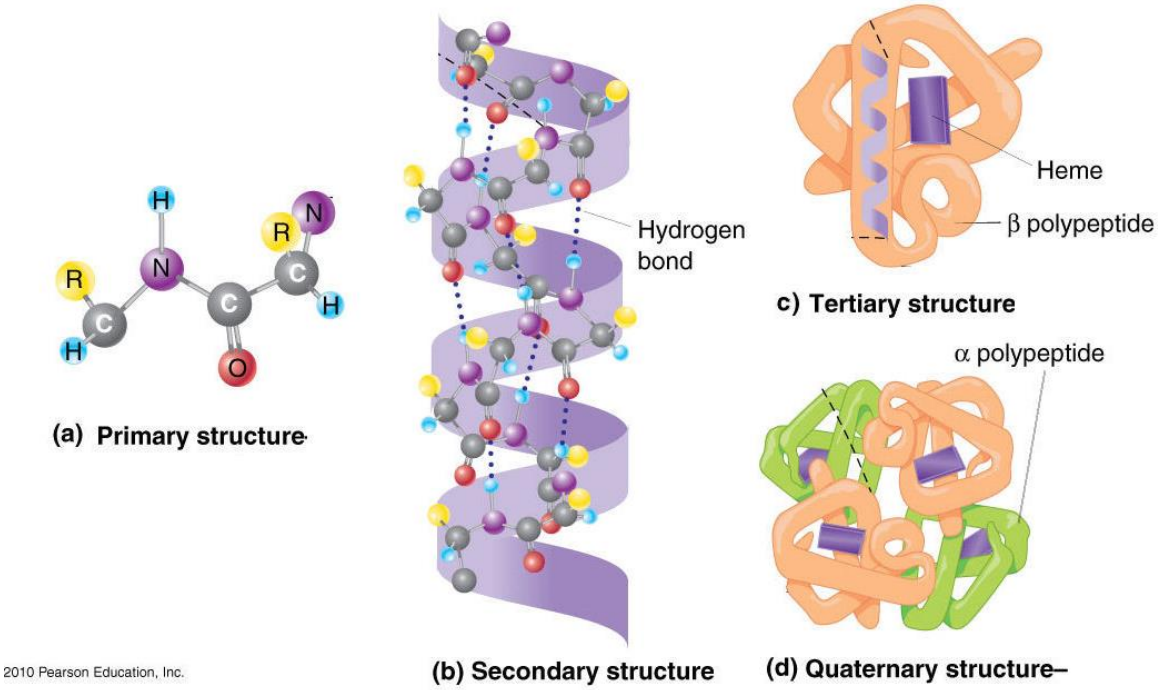

Figure 3 Primary to Quaternary Structure representation (Source: Pearson Education, 2010)

Primary structures consist on the amino acid sequence, with one-dimension information. In secondary structure, three-dimensional information exists, with an indication of how the amino

acids relate between them using high and low strength connections. Tertiary structures are the result of every interaction in a single peptide chain. Many proteins end in this level of complexness. However, others need to be inserted in a protein complex. Quaternary structures are these complexes of several peptide chains, that perform a function in the organism [10].

2.2.1 Primary Structure

Primary structure, as defined previously, have one-dimension data, the amino acid sequence. Being a simple description of a protein, it presents as the backbone of all other features it possesses. All other structures, secondary to quaternary, have the basis in it, and its properties.
As the amino acid sequence is augmented, all the biochemical factors in the set of amino acids already present are influencers to the next amino acid positioning, and how it will shape the secondary structure.
Therefore, the data analyzed in structures other than primary want to summarize the information that primary structure gives, i.e. if a given protein has a pH of 5.7, it is because the conformation and pH's of the primary structure led to this overall result, and not an intrinsic property of the protein.

2.3 Protein function

Each protein, depending on its conformation, is designed to perform a different function in the organism. Function is dependent of a net of factors including:

a) Surrounding Environment

In biological environments, there are certain variables that enable or disable mechanisms. Temperature, concentration and pH are some of the variables that play major roles as constraint to the functions and interactions that protein and their complexes execute in the said environment, and even make them obsolete, as in the case of the denaturation of proteins [10].

b) Internal presence of each amino acid

This factor is the given relation that each amino acid has with the sequence. This factor can be observed in different ways: extreme, relative and positional [11].
The extreme approach is the feed of the number of amino acids, not considering any other factor of the sequence.
In the relative approach, we get a sense of the whole of the sequence and we lower the level of pureness of the factor introducing it as a percentage of the total number of bioproducts.

And in the positional approach, we introduce another important relationship giving position an important role. The existence of 30 samples of lysine in the beginning of the sequence or spread throughout the protein should be considered differently in the application of this approach.
Each approach has its pros and cons, presented latter in feature selection and encoding strategies.

c) Conformation

Conformation is one of the fundamental features for the protein complexes structure determination. It is the shape that a given protein or amino acid sequence takes in a certain environment. The adaptation to these biochemical circumstances results in subtill or enormous changes that will derive in different complexes or functions depending on the environment. This characteristic is one of the biggest arguments against a simplistic approach to a protein, based solely on the amino acid sequence, for the creation of bioinformatic models [12].

d) Protein Complex

The composition of these proteins is mainly defined by their folding and their conformation. The association of chains between the same protein are also important for the definition of the biological activity.
There are several factors that affect these two characteristics, such as pH, concentration of each component, the size of the amino acids, and so on.
A protein complex is, therefore, an association of amino acid chains or proteins that, considering the surrounding environment, have coupled in specific sites creating a single active unit.
This active unit, in most cases, is the true executer or conditioner of a given process in the organism [13].

2.4 Protein-Protein Interaction

In human proteome, based on UniProt [14], there are nearly 70,000 proteins in human organism. Bibliography varies differently from this number, considering or not the influence of splicing in the formation of proteins with different properties. Therefore, the number of proteins ranges from 20,000 to 140,000. The number of combinations of proteins, in these extremes, is 200 million or 9.8 billion, respectively.
For any real number however, it is known that the number is far too large to be tested in laboratory experiments. If we go to a different dimension, of the interspecies domain, UniProt presents 158 million different estimated proteins. This would result in 1.25 E16 combinations.

This term relates to a specific coupling or complex of proteins that complement each other in the development of a function, or in the inhibition of one. This process is conditioned by the surrounding environment, since the coupling of two proteins depends on the conformation they assume. As described before, conformation is relevant to protein interactions and the formation of complexes, since the interactions between certain amino acids can only occur if they are accessible and the point of contact is in the ideal conditions of coupling.

Other factors are of special relevance to the existence or conditioning of an interaction, considered even as regulatory means to the interaction or decoupling of proteins in the respective environments.
For example, high levels of concentration of a protein can induce the interaction, given the necessity to discard that element and generate an equilibrium, or degrade the interaction as a strategy to increase its low concentration, if the opposite situation occurs.
On the other hand, the concentration of other proteins has similar effects. If other proteins or subtracts are present in higher concentrations, then the interaction is more likely to occur [15].

The structure of a protein, with the solvent accessible area and larger or smaller number of ligands is also a characteristic considered as essential [16]. As well as the difference or similarity between the ligands present is a connected influence factor.

Chemical factors are also subject to other interferences, such as covalent modifications in the proteins and amino acids that compose them [17]. These factors are also connected to physical factors, such as the electric fields surrounding the proteins that interact [18].

2.4.1 Hotspot

Hotspots are known interaction zones, characterized by sequences or chemical conditions, derived from specific folding, which have a high probability of attracting other protein interaction spots [19]. These regions are subject to study and modeling despite the heavy computational weight involved in the several steps (from sequencing, to folding, to hotspot recognition) since they have high value in protein detection.

All the terms discussed in the previous segments have a foundational role in PPI detection, especially in computational model development. All of them have been used as a hypothesis for the improvement of feature selection and engineering or differentiating weights for classifiers, as described in the State of the Art.

3 State of the Art for Protein Interaction Detection

The problem of Protein Interaction Detection, and the foundation for this master thesis, has been the object of many approaches. Form biological methods to computational model developments, this chapter is dedicated to finding the best results in the field, focusing on the ones that have a resemblance to Autoencoders structures. The general Protein Interaction Detection problem is described in the workflow in Figure 4.

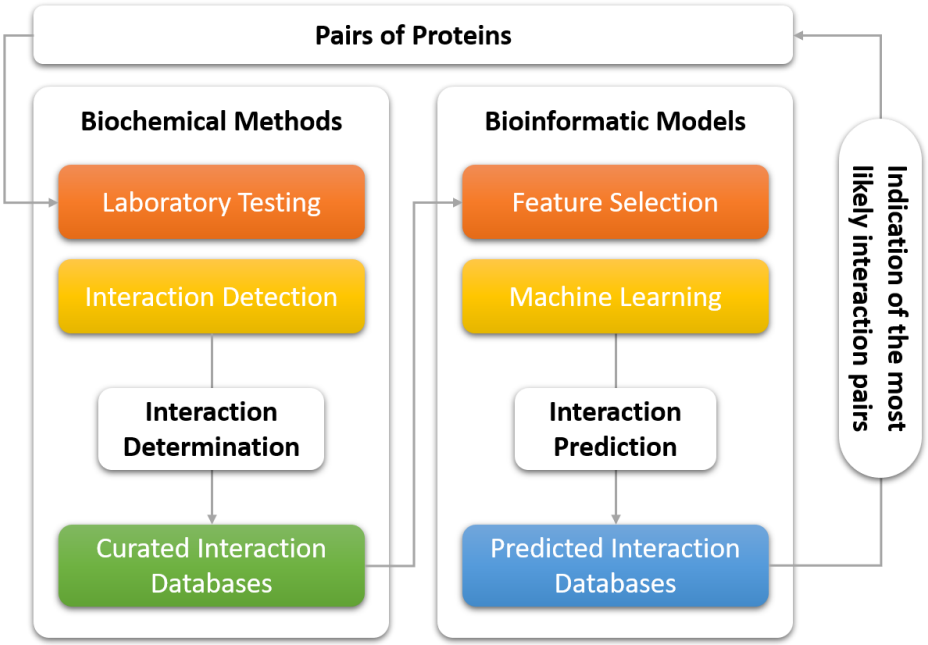

Figure 4 Representation of the Protein Interaction Detection Problem and the Pipeline used for the development of Curated and Predicted Interaction Databases

The Pipeline starts with the establishment of a protein pair to be tested for an interaction. Through one of the possible methods to implement, the Laboratory testing determines the existence of an interaction or not, feeding that information to the curated dataset for its storing and availability for study.

These curated datasets, based on the Biochemical methods, are then applied to the field of Bioinformatics, where the diverse algorithms are trained and tested in order to develop models that can be applied to the full set of interactions and protein pairs that exist in the different proteomes.

The application of these models in such large scale, enables the creation of gigantic databases containing predicted interactomes. These interactomes are, however, uncertified by biochemical methods. The intention is to predict interactions that are tested using those methods, and certified, building more examples to add in the curated datasets.

The first database specialized in Protein Interactions was the Protein Interaction Database, created in 2000. Since then, the number of publications has increased and with them, the number of

databases curated and non-curated [20]. The main literature curated datasets are BIND, BioGRID, DIP, HPRD, IntAct, MINT. These databases consider proteins from several organisms, with the exception of HPRD, exclusively of Human interactions. The fact that they are literature curated implies a similarity between the number of interactions, despite the different requirements, reaching 20 to 30,000 protein samples and between 50 and 100,000 interaction samples.
On the other hand, prediction databases follow the process described, based on the existing curated datasets. Since the models of Protein Interaction prediction is different between databases, the number of interactions have high sparsity, between 200,000 and 88 million samples. The main predicted databases are MiMI, PIPs, OPHID, STRING and UniHI.

3.1 Feature Selection

Incorporating all these specificities, an analogy is conceived to the three main feature selection strategies. While filter methods need no incorporation of learning algorithms to define the subset of features used [21], wrapped and embedded methods have different approaches.

Wrapped feature selection can be divided in forward or backward selection, depending if the set of features is obtained by the addition or removal, respectively, of dimensions dependent on their performance in a basic learning algorithm [22].

The embedded methods consider weights for all features, adapting their weight in each iteration of a performance assessment, with some of these reaching 0, making the feature irrelevant [23].

3.2 Biochemical Methods

In the field of Protein Interaction Detection, only laboratorial methods, known as biochemical methods, present nearly 100% of accuracy. However, these methods are ordinarily expensive and waist a significant part of available resources and raw material.

Bimolecular fluorescence complementation

This method consists in adding two different, but complementary, molecules to the studied proteins. If the molecules interact, the two counterparts of the fluorescent element interact as well, creating a detectable luminescent effect that certifies the existence of interaction [24].

Tandem affinity purification

Based on immunoprecipitation, this strategy uses tags introduced in a given cell or biological environment. The tags create a differential of weight that allow steps of purification through centrifugation. The results of these purifications consist in the target protein and their interactors, which can be latter identified by mass spectrometry or gel electrophoresis [25].

Proximity Ligation Assay

Considering antibodies' reaction to proteins, this method reacts to a connection of proteins by creating a ring with the antibodies with bio probes designed in laboratory. These probes contain DNA and they start a process called rolling-circle amplification, in which DNA polymerases increase the size of the molecule. The detection of the interaction is possible by this effect, given that the resulting molecule increases size until up to 1000 times [26].

Co-Immunoprecipitation

This process follows the premises of the above, creating a connection between a known protein, and its antibody. The antibody sticks to the known molecule, which is interacting with unknown proteins. The full complex is extracted and analyzed through Western blotting, discriminating which proteins are present in the final solution, part of the interactome of the original known protein [27].

Relevance for bioinformatic models

As bioinformatic models are based on the results given by biochemical methods, we need to extrapolate the functioning of the latter to the first.

Common to not only these, but to all biochemical models, is adjacent the idea that the sum of the features is what gives a signal of interaction, whether it is a fluorescence, increased size effect or precipitation of molecules. This statement may sound obvious, but it excludes the extrapolation that a given protein, by itself, may determine which is its interactome. The cause for this unavailability is the fact that any proteins are dependent on the environment in which the interaction occurs to be available or not for interaction.

The second important point is the existence of a third element, traditionally an antibody. In bioinformatic models, the existence of this entity needs to be simulated. This approach is usually not implemented since the models already work with the intended proteins and don't have to isolate them from the crowd of proteins in a given environment.

3.3 Computational Models

Given the necessity for a better selection of proteins to be tested in laboratory, bioinformatic investigators have dedicated a lot of effort in the discovery of new models that can successfully predict interactions based in known features.

Nowadays, a prevalence of machine and deep learning approaches have the most important share of ongoing work in the field, given the recent developments in those areas [28].

Different articles in the Machine Learning branch of bioinformatics have tested different algorithms, favoring Random Forest algorithms, Support Vector Machines, K-Nearest Neighbors, Autoencoders and Ensemble Algorithms.

3.3.1 Random Forest

Random Forest algorithms are developed using decision tree-based processes, where multiple trees are trained and built to acknowledge the problem. Each tree is divided in true and false in each feature, with values and biases that affect the decision and are adapted in every iteration. The majority class voted in the decision trees is the final label attributed to each protein pair. As Figure 5 shows, Random Forest algorithms base their final probability attribution in the satisfaction of several conditions to which the highest correspondent in all decisions is the selected class.

Using these classifiers, Bi-Qing Li et. Al used two and three-dimensional protein features to predict PPI with 78.99% of sensitivity [29], while the prediction of protein to mRNA binding sites by Zhi-Ping Liu et. Al reached 82.4% accuracy [30].

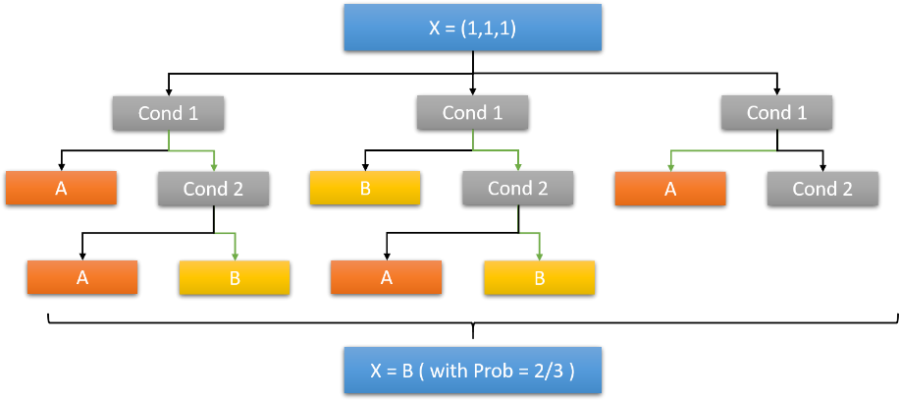

Figure 5 Schematic Representation of Random Forest and the result obtained using probabilities.

3.3.2 Artificial Neural Networks

These networks have the possibility to create unlimited layers of functions that transform a given input feature space in a single output with the desired outcome. In the case at study, the final layer should be a one-dimensional layer, with 0 and 1 corresponding to the prediction of interaction or not. This layer is the result of the connections between intermediate layers and functions learned in the training stages and followed by the test datasets.

As we can see in Figure 6, the input can be of a dimension that does not comply with the hidden layers of the Neural Networks, since the layers between the input and the output need to store information and functions that are developed in the training of the algorithm, in each cell.

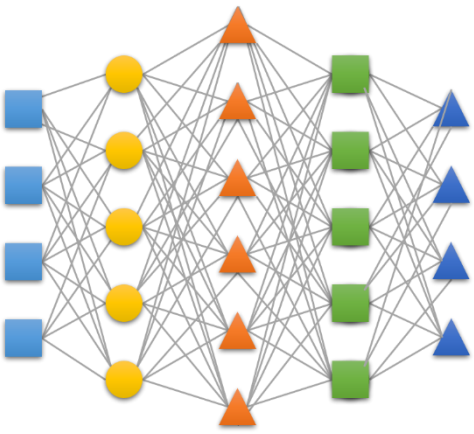

Figure 6 Artificial Neural Network General Structure, with the Hidden Layers it contains. The blue squares represent the input and the output is represented by the blue triangles.

In these networks, several methods including autoencoders have been used, achieving state of the art results year after year.

The use of the Discrete Cosine Transform is also a development, coupled with an SVM and k-NN classifiers, achieving performances of 84% in AUC metrics [31].
In the deployment of Position Specific Scoring Matrices, other works have used Principal Component Analysis and Evolutionary Networks, obtaining 94.1% and 80% accuracy in testing datasets, respectively [11].

3.3.3 K – Nearest Neighbors

The Classification problem in the context of protein interactions possesses two classes. This means that, if a separation surface exists, two neighbors must always be from the same class, with exception of the surface itself. Nearest Neighbors Algorithms take this process and apply it to a *k* number of neighbors, i.e. if k = 5, and the five nearest neighbors have a majority of class A, then the point considered will also be labeled as A.
In the context of protein interactions, it was used by Liang Lan using multi protein data sources to train and classify protein interactions based on their characteristics, achieving 84.8% of AUC [32]. It was also used in few thousand interactions by Mario Guarracino, achieving 98.11% with 1-NN classifier [33].

3.3.4 Support Vector Machines

As classification algorithms, Support Vector Machines are widely used in the area of Machine Learning. The goal of these Machines is to create hyperplanes that manage to separate classes with maximum distance between them, resorting to the perpendicular vectors to the distance between dots of different classes.
This method has been used in the field of PPI, associated with evolutionary properties, to improve the accuracy to 90.57% accuracy [34].

3.3.5 Ensemble Algorithms

In Protein-Protein Interaction detection, several algorithms also try to ensemble different approaches. The most common techniques involve the combination of Support Vector Machines as a classifier, after the use of an algorithm as a first filter of the features. Works such as "*Prediction of protein-protein interactions from amino acid sequences with ensemble extreme learning machines and principal component analysis*" show the complement between two different structures, and not the mainly used for this problem, with an accuracy of 87.50% [35].
Stacked Autoencoders have also been used with SVM ensemble, with 97.07% area under the curve.

Other works used coupling strategies between Linear Discriminant Analysis and Random Forest, Ensemble Autoencoders, Simple Classifiers with Random Forests. All these structures have achieved result in the magnitude of 80% at least [36].

3.4 Autoencoder Methods

The central approach of this thesis is focused on autoencoders, machine learning structures that try to encode a certain input in one or several hidden layers of reduced dimension (in most cases, being denoising the exception), and then decode it with the maximum level of confidence, reducing the error of the final reconstruction, as demonstrated by Figure 7 and its equation.
The reduction demonstrated in the figure is one of the key elements of Autoencoders. In the lower dimension Hidden Representation, this Neural Network is capable of storing the necessary information and then reconstruct it, generating X' from X.

$$input \longrightarrow HL = encoding(input) \longrightarrow output, error = decoding(HL, input)$$

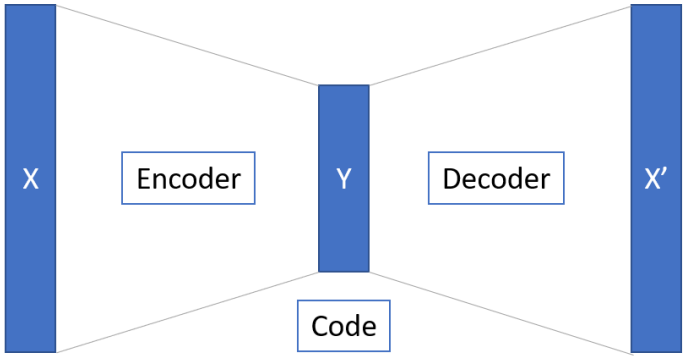

Figure 7 Simple representation of a standard Autoencoder structure, with the input and output (X and X') and the Hidden Representation (Y).

Autoencoders have many functions for the encoding and decoding of the input and the hidden layer(s) allowing for a search of parameters. Besides, different structures of autoencoders exist, as mentioned below. If a comprehensive autoencoder structure is found for feature encoding, we can work with the Hidden Representation to feed a classifier of our choosing.

3.4.1 Denoising Autoencoders

Denoising autoencoders work with corrupted inputs and try to reconstruct the original input based on similar inputs that have been encoded and decoded. The objective function to be reduced is the error between a corrupted and an uncorrupted input.

In Figure 8, a representation of this type of autoencoders show that a corrupted input in $\dot{x}$ can be interpreted and completed with the use of the network built with the full feature representation.

This type of Autoencoders presented promising results in the field of image de-noising, and in broader terms, in all the processes of increase the robustness of features [37].

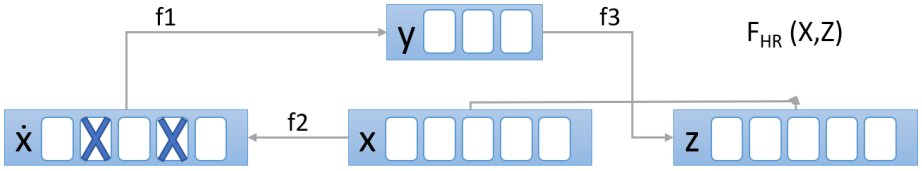

Figure 8 Representation of a Denoising Autoencoder, where ẋ is the corrupted representation.

3.4.2 Autoencoder Ensemble

Some works have used ensemble of structures, including autoencoders. This assembling may be between autoencoders or autoencoders and other networks, such as Extreme Learning Machines, or Support Vector Regression. These strategies have showed promising results in the field of time series analysis [38].

As Figure 9 shows, the autoencoders are lined-up in a unique-level ensemble to which each autoencoder attributes a label.

In the case of Autoencoder ensemble, the strategy may be named as a Stacked Autoencoder, if the input of the ensemble is the previous Hidden Representation obtained.

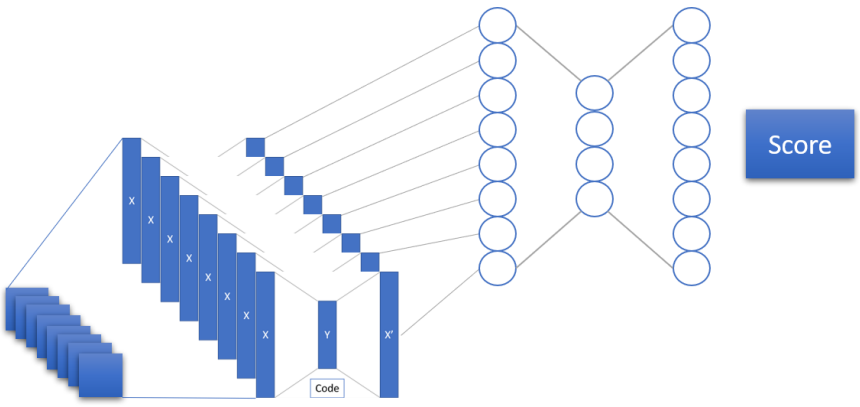

Figure 9 Representation of an Ensemble Autoencoder.

3.4.3 Variational Autoencoders

In Bioinformatics, variational autoencoders have also been used for our proposed applications, but with different approaches. The variational component is added in the stages signaled in orange, the stage that interpret the responses as mean and standard deviation (respectively down and up in Figure 10.

Cancer drug targeting has benefited from research using these autoencoders that instead of fixed feature values, learn the distribution of values, and return a mean and standard deviation values in which the models are then based [39].

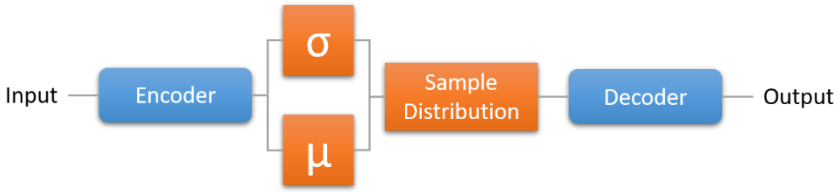

Figure 10 Representation of a Variational Autoencoder

3.4.4 Stacked Autoencoders

There are also exploratory works in the field of sequence-based Protein-Protein Interaction that use Stacked Autoencoders, structures that consist in more than one hidden-layer, building several encoding functions to reach a maximum stretched representation.

In Figure 11 the sequence of Hidden Representations is visible, reaching maximum efficacy in the Z field. This type of Neural Network implementation needs to pay special attention to the error propagation in the prediction of the several autoencoders

The results obtained in Protein-Protein Interactions showed results in line with the current state of the art on this field [40].

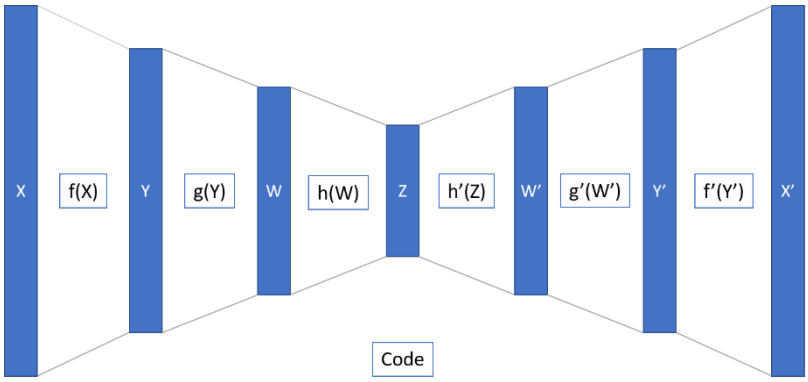

Figure 11 Representation of a Stacked Autoencoder

3.5 Protein Representation and Classification Strategies

In the development of computational models, each structure is correspondent to a biological component or set of components, implicating a correlation in which both biologists, informatics and bioinformatics make a direct relation to what a system or component means in each model. This way, a biologist can correct a computer scientist in the model it composes or understanding what he needs to do differently in his work to achieve better performances with the help of the developed and tested models.

Considering the general combined knowledge of each area, the following correspondences can be established:

a) Sequence representation and encoding

Sequence representation is the baseline of the work, representing a protein. The encoding method chosen, which ever it is, represents a single protein, in its individual state, and not considering any other protein in the process.

b) Pair Label

The label of each given pair is the equivalent to bind or does not bind, attributing 1 to the first and 0 to the latter. In biological applications, this interaction would be detected using measurements of environmental change or by a fluorescent indicator.

c) Concatenation of encoding methods

The result obtained from the concatenation of matrices intends to represent the protein pair as one. Although the sum of properties had also been considered, it would lead to the loss of information regarding the individual properties of each protein, and this sum of properties, if relevant, can still be found in Dimension Reduction or Classifier search.
The concatenation of encoding methods also allows for the exploration of the properties of each protein individually.

d) Dimension Reduction

The use of dimension reduction techniques, such as Autoencoders, is involved in the representation of proteins or protein pairs. Dimension reduction can be compared to the use of different lenses in the observation of these characteristics. A more direct or indirect approach doesn't change the

protein, but the properties we can see in it, or the degree in which we see them. The choice of protein characteristics or combinations is a natural process in the field of biology, that takes several forms in the field of bioinformatics, such as dimension reduction.
It can also be compared to the summarization of features, as described in the Primary Structure Subtitle (see Section 2.2.1), that the secondary to quaternary structure analysis give to the biologist and bioinformatics.

e) Classification models training

The environment in which proteins are synthesized and interact is complex, and hard to describe in chemical factors, as so many have influence, such as percentage of water, temperature and pH. These factors combined will be the description of the environment in various forms. In simplistic terms, a correspondence can be found between each Neural Network or Computational Model function, and a given environment factor.
The creation of models to each component of a single organisms would lead to lack of information, increased complexity and contexts impossible to analyze given the number of proteins in some environments.
Therefore, given the proteins that interact in certain macro environments (in this case, organisms), the intention is to predict which environment exists in the said organism. This is the biological equivalent to training a classification model, in which the classification model represents the organism itself (and its biochemical factors).

f) Classification Model Testing

This stage is artificial, representing a biological simulation. Its correspondence is the creation of man-made environments, considering the models built in Training, and use protein pairs already known to interact or not to test if the environment is the same, giving validity to the biochemical factors found.
In Bioinformatics, it is the validation of the functions of the model trained, which are applied to already known pairs that bind or do not bind.

g) New pairs test

After the classification model testing, random or directed (conditioned) tries are made to test new pairs of proteins that will interact in the context of the trained model. This is the bioinformatic equivalent to the experimentation, in the same biological environment (organism), of new protein pairs.

h) Interspecies Benchmark

We know that many biochemical factors are similar between species and, depending on the level of correlation of two species, a pair given to two organisms can have the same result (bind or does not bind). This stage intends to simulate, not only the same pairs, but different pairs. This means that we take a trained model (an organism) to which several features (protein pairs) are fed and the result can be the same or different, giving information to the potential in each organism. Or even achieving results of proteins never found in the same organism, opening doors to proteins of some organisms interact with proteins of other organisms, in the latter.

4 Datasets and Methods

This chapter focuses on the processes used to go from the original datasets, HitPredict and BioGrid, to the result of the classifiers, describing the methods used in each step, the inputs and outputs, with respect for the correspondence made in the previous Chapter.

4.1 Software

The methods and datasets were implemented and processed, respectively, using MATLAB [42], a computing environment developed by MathWorks specialized in numeric calculations, specifically matrixes, hence the prefix MAT.
In the analysis of the final tables and datasets as well as small calculations, software Microsoft Excel was also used, as well as in the creation of graphics and colored tables.

4.2 Datasets

4.2.1 HitPredict dataset

The original dataset is composed with the ID's of the human family of UniProt. Taking those ID's, we programmatically access to the HitPredict Database, with nearly 17,000 protein examples being inserted in the database [43].
From the HitPredict Database, we extract, for each protein, an Interaction Matrix, described as followed in Table 1.

Inter	ID UP	Expe.	Method	Score	Annotation	Sc Int	Confidence
558263	Q9UQL6	HDAC5	Small-scale	0.46	0.60	0.527	High
484618	P52952	NKX25	Small-scale	0.42	0.60	0.501	High
290816	B4DJ51	B4DJ51	High-throughput	0.44	0.50	0.470	High
547134	Q9GZU8	F192A	High-throughput	0.38	0.55	0.458	High
145728	P61289	PSME3	High-throughput	0.38	0.23	0.258	Low

Table 1 Example of Interaction Matrixes obtained from HitPredict

Each interaction is identified by the field Inter, a cut for the word Interaction, and makes a differentiation for every interaction at Hit Predict Platform, identified or curated by the Experience identified in the "Expe" field, with a large or small-scale level experience (small-scale vs high-throughout) and a confidence level attributed in the Score field. The UniProt ID of the second protein (the one that the original matrix protein interacts with) is present at the ID UP.

The Score Interaction (second to last column) is an attribute developed by the HitPredict authors that consists in attributing an higher coefficient to more recognized experimental methods of determination of interaction or higher number of experiences.
The threshold for the definition of an interaction as of "High Confidence" is 0.281. Below that, the interaction is considered of "Low Confidence", as described by the article in which the platform is based on [43].
In the selection of pairs for the final dataset, we divided the set in the negative and positive approaches.

Positive approach
The positive set of pairs is formed using the following steps:

1. Selection of a protein;
2. Interaction Matrix of the protein using the HitPredict programmatic access;
3. Selection of the pair with the highest Interaction Score, verifying that it is above the 0.281 threshold

Negative approach
The negative set of pairs is formed using the following steps:

1. Selection of a protein;
2. Interaction Matrix of the protein using the HitPredict programmatic access;
3. Selection of another random protein ID;
4. Verification that the new protein is not present in the Interaction Matrix.

Final HitPredict dataset
The final dataset consists in a balanced dataset of 16,672 pairs of protein ID's, 8,336 with positive approach and 8,336 negative approaches. With the ID's chosen, we retrieve the sequence from the UniProt Database, and store it in a structure.
This dataset is the base for the encoding options that create three different datasets based on the Encoding methods.

4.2.2 BioGrid dataset

The BioGrid datasets is based on the BioGrid Downloadable dataset [44]. In this dataset, each interaction is described in the form of a 1x3 array with the UniProt taxonomy identifier, and two UniProt Protein ID's [14]. The dataset contains only positive interactions, and no Negatome was used in the research.

The total number of processed organisms present in the dataset was 52, and it accounted for 90,697 different proteins and 2,008,893 Protein-Protein Positive Interactions, of the 8.2 billion possible (from the cross interaction of all proteins). From those 52 organisms, were excluded the ones with lowest amounts of data, generating a high threshold of 30,000 interactions to limit the processing, and a lower threshold of 35 to ensure that the data was significant. The one exception was the Human dataset, in which more proteins were selected, since it is the one with most interaction, more interest to the clinical studies, and to which we want to compare all other datasets. Figure 12 shows the size of all the datasets used, in which the lower threshold of 35 can also be observed. In order to increase its interpretation, values in the vertical axis are represented in a base 10.

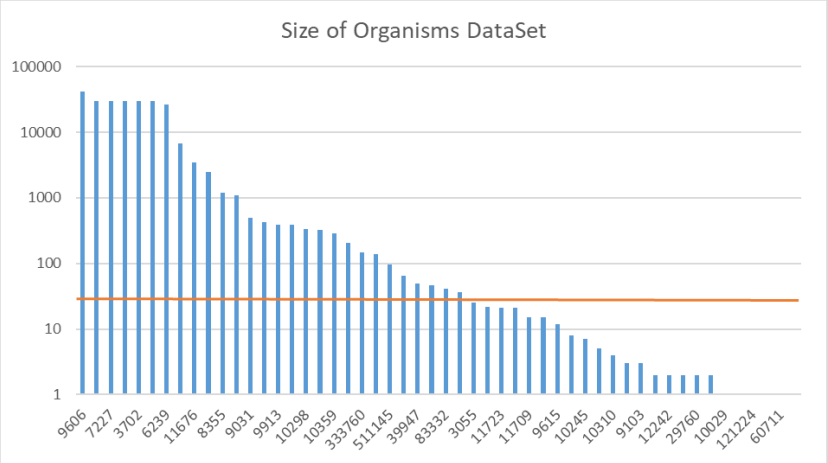

Figure 12 Organisms by number of positive interactions and the red bottom size limit of 35 positive interactions per dataset.

The final dataset was composed by 26 organisms presented in a table in Annex 3, with the number of positive interactions for each considered.

4.2.3 Preprocessing and Negative Pairs Generation

Each dataset consisted in the definition of a structure in which the positive pairs in the datasets were generated by the extraction of both UniProt sequences, based on the UniProt ID's in the original BioGrid dataset.

The negative pairs were generated using two random indexes. The first index extracted an ID from the original dataset, and the second ID would be verified not to be an interaction with the first chosen protein. After that, both UniProt sequences would be extracted in the same procedure described before.

4.2.4 UniProt Programmatic Access

UniProt is the Universal Protein Database [14]. It is referenced as one of the top leading protein banks, supporting validity to the data produced and stored in it, with high levels of curated datasets. To access the database, a command was used to receive the FASTA file format, through the use of *urlread* function, and the ID's obtained in each of the previous datasets described. After the reception and decryption of the FASTA File, the preprocessing and encoding methods were applied [45].

FASTA format is a general descriptor of a protein or DNA/RNA, used to store basic characteristics of amino acid or nucleic acid chains, such as the name of the protein or gene, and the sequence. It starts by the character '>' and is followed by the main characteristics. In the line below, the sequence is inserted resorting to the amino acid single-letter code (Annex 4).

4.2.5 Overlapping between datasets

Several protein interactions are characterized by their presence in several organisms, being a measure of the similarity between organisms and their complexity. A small study was conducted between all the 26 studied datasets of interactions, representing the overlapping between them.

Since protein ID's are different for each protein in different animals, the search was conducted through the sequence and not through the protein ID.

The results are presented in Figure 13, which is also present in the Annex 5 with higher precision values.

	10090	10116	10298	10359	10376	11676	227321	237561	284812	333760	3702	37296	3847	39947	511145	559292	6239	7227	7955	83332	8355	9031	9606	9823	9913	9986
10090	1,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,001	0,000	0,000	0,000
10116	0,001	1,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
10298	0,000	0,000	1,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
10359	0,000	0,000	0,000	1,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
10376	0,000	0,000	0,000	0,000	1,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
11676	0,000	0,000	0,000	0,000	0,000	1,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
227321	0,000	0,000	0,000	0,000	0,000	0,000	1,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
237561	0,000	0,000	0,000	0,000	0,000	0,000	0,000	1,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
284812	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	1,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
333760	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	1,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
3702	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	1,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
37296	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	1,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
3847	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	1,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
39947	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	1,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
511145	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	1,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
559292	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	1,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
6239	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	1,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
7227	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	1,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
7955	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	1,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
83332	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	1,000	0,000	0,000	0,000	0,000	0,000	0,000
8355	0,000	0,002	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	1,000	0,000	0,000	0,000	0,000	0,000
9031	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	1,000	0,000	0,000	0,000	0,000
9606	0,001	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	1,000	0,000	0,000	0,000
9823	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	1,000	0,000	0,000
9913	0,008	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,005	0,000	1,000	0,000
9986	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	1,000

Figure 13 Overlapping matrix between organisms' datasets.

Figure 13 shows all the 26 datasets correspondence, where green means that the overlapping is total (every dataset contains itself entirely), yellow means that the overlapping is different from 0%, and red means the overlap is nonexistent. No overlapping result (in yellow) is over 0.77%, reason to why we can consider that the datasets are independent, and don't interfere with the final results obtained.

Although some datasets have considerable overlap in the overall datasets, the fact that the most complex organisms are truncated to 30 000 interactions, and the interactions are randomly chosen across the dataset, makes the overlap difficult to occur. Therefore, we can present here datasets that do not overlap in organisms that have phylogenetic relationship, but to which the common interactions were not considered. This evidence shows that the datasets meet the criteria for the experiments.

4.3 Feature Engineering

The definition of a point in space or the representation of an object require the use of features. The quantity and quality of the features used are of key importance to the classifier later applied. A feature must always be measurable, in order to be understood by a classifier or neural network and should be as independent as possible.

Measurability is the opposite of the intended effect with dimension reduction, in which all the focus shifts to the quantity of features, their quality and independency, specially the latter. If any

dependency is found by a dimension reduction algorithm, it is aggregated in a single feature, increasing its quality, and decreasing its measurability and understandability.

The specificity of the selection of the features leads to thorough studies of the existing properties and which should be the main features used in the work developed.

The approach used in feature selection can be holistic or individual. Both have positive and negative points which should be analyzed, and then considered, taking the characteristics described before in account.

4.3.1 Protein related features

Features related to a holistic approach of the proteins have a deeper understanding of the environment that surrounds a protein, considering the set of amino acids that are part of the sequence and the biochemical properties as a sum.

The main properties considered were:

a) Amino Acid Composition - composed by 20 features, in which every value is the count of all numbers [46];
b) Dipeptide Composition – composed by 400 features, in which every pair of two followed amino acids are counted [47];
c) Conjoint Triad – composed by 343 features, representing a head count of all the triplets existing after the encoding according to specific groups of amino acids (see Annex 2), which are divided in 7 categories [48];
d) Autocovariance – composed by a matrix calculated by a variable number of properties, in which time series are taken in consideration in the evaluation of the original features [49];
e) N-grams – composed by different number of features depending on the value to n, but to which the structure is similar to Dipeptide and Conjoint Triad if n = 2 or 3, respectively [50];
f) Motifs – pieces of protein information encoded by other strategies or in textual input, that characterize strong ligand zones in the proteins [21];

Analyzing these features, we can capture the holistic view of each protein, with the surplus, in some cases, that the properties of the amino acids are part of the formula or count made by the method. However, a holistic view misses some of the individual characteristics of the amino acids still, and many times releases the components from the sequential nature, making general sums of the full quantity of the amino acids presents, considering or not their properties in the overall.

4.3.2 Amino acids related features

Regarding the individual features of amino acids, we could think of a near infinite number of properties. Since they are analyzed individually and considering the study of amino acids has been thorough regarding their importance in protein functions, and life in general, all its properties can be measured with high precision and each of them could be a feature. The general table of chemical properties for each amino acid is described in Annex 1.

The main physical properties of the amino acids are:

- Net Charge Index of Side Chains
- Volume of side chain

While the main chemical properties are:

- Polarity
- Polarizability
- Hydrophobicity
- Hydrophilicity
- Solvent accessible surface area

4.4 Encoding

For the encoding of the dataset, three different strategies were used, based on two different encoding methods.

4.4.1 Conjoint Triad

Conjoint Triad Method was created in 2007 [48] with the intention of describing proteins in a simple descriptive and fixed array, in which both holistic and specific amino acid characteristics have weight in the classifier chosen.

The Conjoint Triad method allows for the identification of trios of amino acids with certain properties, according to previously designed tables, in annexes Y.

The previous 7 labels are then attributed to each sequence, creating an integer sequence. We then proceed to create groups of three amino acids, with an advance of 1. We create a 7x7x7 matrix, with a “xyz” coordinate for each trio.
The matrix is then introduced in a structure for each matrix.

4.4.2 Autocovariance

Autocovariance is a statistical measure introduce in the early 1900's in which the values of a vector can be assessed in a time-series approach [49].
Using the formula explained in the article, a 7x30 matrix for each protein sequence, the final matrix is calculated for that is then introduced in a structure for storage. The formula is:

$$AC_{lag,j} = \frac{1}{n-lag}\sum_{i=1}^{n-lag}\left(X_{i,j} - \sum_{i=1}^{n-lag}(X_{i,j})\right)\times\left(X_{i+lag,j} - \sum_{i=1}^{n}(X_{i,j})\right)$$

- Lag is an arbitrary number here defined as 30, since it is the number used in previous works
- X is the vector of properties
- Properties are defined by the table in annex 1.

4.4.3 Mixed Encoding

The Mixed Encoding strategy involves the concatenation of both previous encoding methods, creating a 553-long sequence of doubles.

4.5 Autoencoder Approach

Our main approach is built on the Autoencoder Hidden Representation [42]. These structures intend to reduce the dimension of a given dataset, without losing capability of regenerating the input. Depending on their structure, several approaches using these networks have been used in bioinformatics and other applications.

4.5.1 Contextualization of Autoencoders

During the development of the thesis, a study was made, in the context of Autoencoder's structure development and familiarization, in which several parameters were studied. The paper sustained the idea that Autoencoders can achieve high accuracies in the prediction of Protein functions (and families).
The use of Autoencoder function have the obligation to establish a Hidden Layer size. A fixed number of hidden sizes would ease the evaluation of a classifier but create other problems such as

approaches in which the hidden size would be higher than the original dimension. Therefore, a percentage-based approach was taken, and the best results were achieved using a 10% Hidden Representation related to the original input size, reason to why this will be the reference value for the use of these structures [41].

4.5.2 Encoding and Decoding Transfer Function combinations

In the choice of parameters in the Autoencoder structure, two main points create the deepest differences of programming and in the outcome generation.
The encoding transfer function and decoding transfer function, considering the hyperparameters that may be optimized, are what defines the learning process of the algorithm. Therefore, their understanding is critical in achieving a good performance.

Logistic Sigmoid Transfer Function - LogSig

$$y = \frac{1}{1 + e^{-x}}$$

Saturating Linear Transfer Function – SatLin

$$y = \begin{cases} -1, & if\ x < -1 \\ x, & if\ x \geq -1\ e\ x \leq 1 \\ 1, & if\ x > 1 \end{cases}$$

4.5.3 Use of Negative and Positive Autoencoders

The use of Autoencoders presupposes the ability by it of the representation of an entity with fewer characteristics than the original ones, achieving dimension reduction by a two-function determination (encoding and decoding transfer functions). Therefore, the Autoencoder is a set of function that can attribute characteristics of an entity to a certain point, with a margin of error. The biggest the margin of error, less likely it is that the autoencoder is an appropriate function to represent the point intended.
In this work, a system was developed where two different autoencoders are trained for each combination of Encoding and Decoding Transfer Functions,

4.5.4 Errors and Hidden Representations as two different sub-methods

The analysis of an autoencoder leads to different possible interpretations of its Hidden Representation. The widely used approach is the Hidden Representation prediction using the model trained, where it is applied the following logic sequence:

a) The reference point is the origin of the Autoencoder, i.e., is the point where the Autoencoder predicts the medium point;
b) Applying the model, all points around the reference point are categorized in each of its features;
c) The final features (reduced dimension) of each point are now the new features to be classified in later stages.

However, using the errors of a Hidden Representation, and as the classifier input, the logic sequence is different:

a) The reference points are the origins of each Autoencoder;
b) The point at study is measured, but the only feature to be stored is the distance to the origin (error) of each Autoencoder;
c) The distances to all autoencoders (errors) trying to find my features are the features to classify.

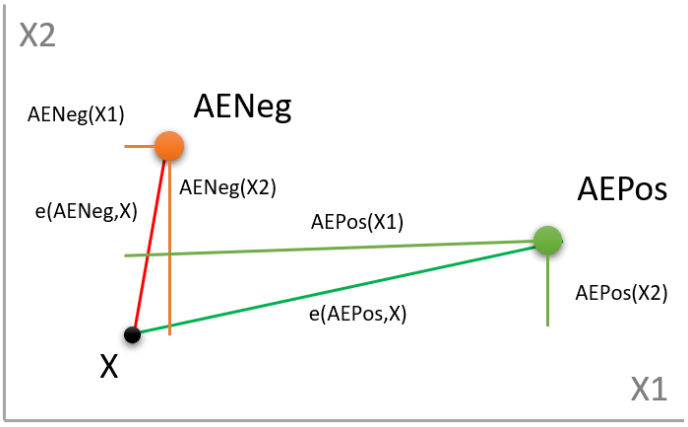

Figure 14 Difference between Error and Hidden Representation Sub-methods.

The logic sequence in both approaches is represented in Figure 14, where the difference can be seen, in two dimensions, between the use of a Positive (AEPos) and a Negative (AENeg) Autoencoder to predict all the features, or just the error (function e) in predicting the point X to be analyzed.

The final result of these approaches is dimensionally different as the use of Autoencoders to obtain the errors makes the number of features equal to the number of Autoencoders used and the representation in new coordinates increases that number to the product between the number of Autoencoders and the dimensions of the Hidden Representation.
The final dimension in this figure is 2 in the error approach and 4 in the Hidden Representation. But in our case study, with 8 autoencoders and their Hidden Representation choice, the difference is from 8 to 68, respectively.

4.6 Support Vector Machine Approach

Our main studied classifiers are Support Vector Machines. They are dependent of certain parameters which can be predetermined or searched in grid values. The main introduced specifications are Kernel Functions and Kernel Values.

4.6.1 Kernel Functions and Values

Kernel functions are the ones that establish the shape of the hyperplane that separates classes. The used Kernel functions are linear, gaussian and polynomial.

Linear kernel determines that the margin must be predetermined as a single valued angle in relation to the origin axis.

$$k(\overrightarrow{x_i}, \overrightarrow{x_j}) = (\overrightarrow{x_i} \cdot \overrightarrow{x_j})^d$$

Gaussian kernel is an exponential function determined margin:

$$k(\overrightarrow{x_i}, \overrightarrow{x_j}) = \exp\left(-\gamma \left\| \overrightarrow{x_i} - \overrightarrow{x_j} \right\|^2\right),$$

where the distance is multiplied by the box constraint value (γ), in an exponential function.

Polynomial kernels, on another approach, determine the vectors through the use of a polynomial order to exponentiate the dot product to. The use of d = 1 is the equivalent to the use of linear kernel.

$$k(\overrightarrow{x_i}, \overrightarrow{x_j}) = (\overrightarrow{x_i} \cdot \overrightarrow{x_j})^d$$

4.6.2 Hyperparameter Optimization

To assist the use of the parameters to each model, the values can be search in grid parameters, to obtain lower values of the objective function determined, which represents higher values of AUC.

4.7 Long Short-Term Memory Networks

Long Short-Term Memory Networks [51] are recurrent neural networks built to face time-series modeling improving the results of other RNN's in that specific field.
LSTM contain memory blocks where the model is learned, taking the input and output gates dynamics of multiplication and their errors to a certain number of inputs behind, creating a forget gate that doesn't allow for the backpropagation to go on until the beginning of the sequence.

The properties from Long Short Term Memory Networks make them sequence adaptable. However, since the size of every sequences are different, we need to perform additional preprocessing in these types of Network inputs.
In order to lose less information, or having to add the fewer empty characters possible, the first step is to order the sequences by size and padding the sequences into mini batches to which the sequences have the same tailored size. The final preprocessing required is to split sequences that overshoot the batches.
As Figure 15 represents, the five compositions of LSTM that follow have different functions. The sequence input deals with the padded sequences and introduces them to the LSTM in its core. The Fully Connected layers ensure that every neuron in the LSTM layer is connected, ensuring all final states obtained in the memory blocks are considered in the final classification.

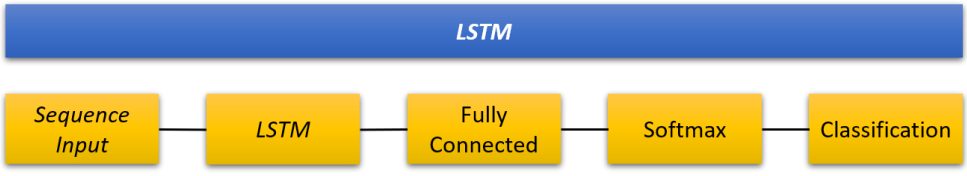

Figure 15 Long Short-Term Memory Network Layers and the hidden stages it represents in the calculation of the final LSTM Algorithm

The development of Autoencoders, being the focus of this thesis, was considered as a hypothesis to be compared to other methods, such as LSTM, which was studied in the following approaches:

a) Train Sequence (Sequence to Sequence) to recognize category

Sequence to Sequence Long Short-Term Memory Networks classify the input sequence regarding sequence labels. These sequence labels are categories or sequences of classes that LSTM predict. These sequences have an important role in automatic translation algorithms.

b) Train Sequence (Sequence to Sequence) to recognize hotspots

As described above, training sequences may imply the result of a sequence not described before, and not pre-labeled. These unsupervised learning machines may present new solutions to previously unknown sequences.

c) Train Sequence (Sequence to Label) to recognize interactions

This is the most direct type of LSTM to the problem approached. Using raw text inputs, we classify the sequences using the LSTM function at MATLAB. The output labels are 0 or 1, according to the training of the datasets.

4.8 Metrics

The final challenge in the organization of the experiments and selection of methods to apply was the choice of the metrics to use.
In the papers read as preparation of the Thesis, and in the general literature, several metrics were used, including Binary Confusion Matrixes (with Sensitivity and Sensibility), F1-Score, Area Under the ROC Curve (AUC) and Accuracy.

Confusion Matrixes are used to evaluate the number of true and false class labeling using a certain model. They can be multidimensional or binary, according to the number of classes used in the problem. Since Protein Interaction is a binary problem, the used notation is: True Positives (TP), the case of a positive interaction being labeled as positive; False Positives (FP), when a negative

interaction is labeled as positive; True Negatives (TN), when a negative interaction is labeled as negative; and False Negative (FN), when a positive interaction is labeled as negative.

Metric	***Equation based on Confusion Matrixes***
Sensitivity	$\frac{TP}{TP + FN}$
Specificity	$\frac{TN}{TN + FP}$
F1-Score	$\frac{2 \times TP}{2 \times TP + FP + FN}$
Accuracy	$\frac{TP + TN}{TP + FN + TN + FP}$

Table 2 Equations of the metrics analyzed as possible candidates

Area Under the ROC Curve (AUC)

This metric is a performance measure which integrates the Receiver Operating Characteristics Curve, the graphic between True Positive Ratio (sensitivity) and False Positive Ratio (1 – specificity).

To reach the best possible combination of metrics, their equations were analyzed considering the following prerequisites:

1. The highest number of True Positives and True Negatives is desirable

The best possible outcome is that the prediction of proteins is of the highest accuracy possible, allowing for the lowest quantity of tests needed, and to build confidence on the model.

2. The lowest number of False Positives is desirable

False Positives and False Negatives in Drug and Protein Interaction Models

Drug trials, and protein interactions trials need to be conservative in the amount of resources used. Though the desired outcome is to predict with the highest accuracy possible the highest number of predictions, the presence of this principle obliges to the importance of reducing False Positives in the preparation of trials. In the building of the model, this is only a fact since the datasets used are balanced, halving each of the true and positive labels.

5 Experiments and Results

As explained in Chapter 2. Protein-Protein Interaction Problem, the intended model to find is the equivalent to the *interactionPredictor*, in the equation

$$y = interactionPredictor(features(p1), features(p2))$$

The planning of the experiments respects this initial equation, describing several steps to be followed, to achieve the best possible model. Since the model was, from the very beginning, thought as an ensemble, several experiments were performed to decide which optimized model could be built with the sequence:

1. Encoding Method Experiment EME
2. Dimension Reduction Experiment DRE
3. Classifier Training Experiment CTE

The convolution of these three experiments translated in the final obtained Pipeline is the equivalent to the following equation:

$$y = CTE\left(DRE\left(EME\left(features(p1), features(p2)\right)\right)\right), \qquad \text{with y = \{0,1\}}$$

This Chapter addresses all the experiments performed as well as their findings, explaining the implications they have in the experiments developed later in the Pipeline.
This equation, and the model predicted, is then applied to the final benchmark dataset, BioGrid.

5.1 Final Experiments Schematics

The set of functions needs to be chained in a general scheme, such as the described in Figure 16, in order to organize the results and the best function in each experiments stage.

In the development of the four experiments, we need to extract the best result for each, without losing the ability to adapt to new organisms the methods we develop. Since we have two datasets, in which we have two different goals, we need to design a schematic for the HitPredict dataset (Figure 16) and another for the Interspecies & Benchmark Experiment (Figure 17).

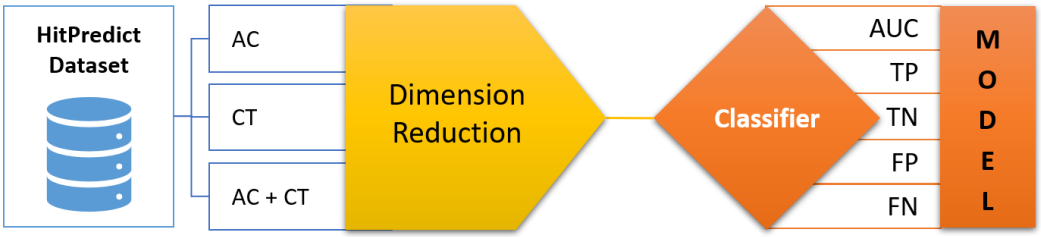

Figure 16 HitPredict dataset Experiment

In Figure 16, we see that the HitPredict dataset is the first step of the experiments, with reduced application, but of exploratory nature in Encoding, Dimension Reduction and Classifier Stages. The main goal of this dataset is to provide viable information of the first two factors, enabling the findings related to the Classifiers in the General Experiment of the BioGrid dataset.

The experiments involved are the Encoding Method Experiment, Dimension Reduction Experiment and Classifier Training Experiment.

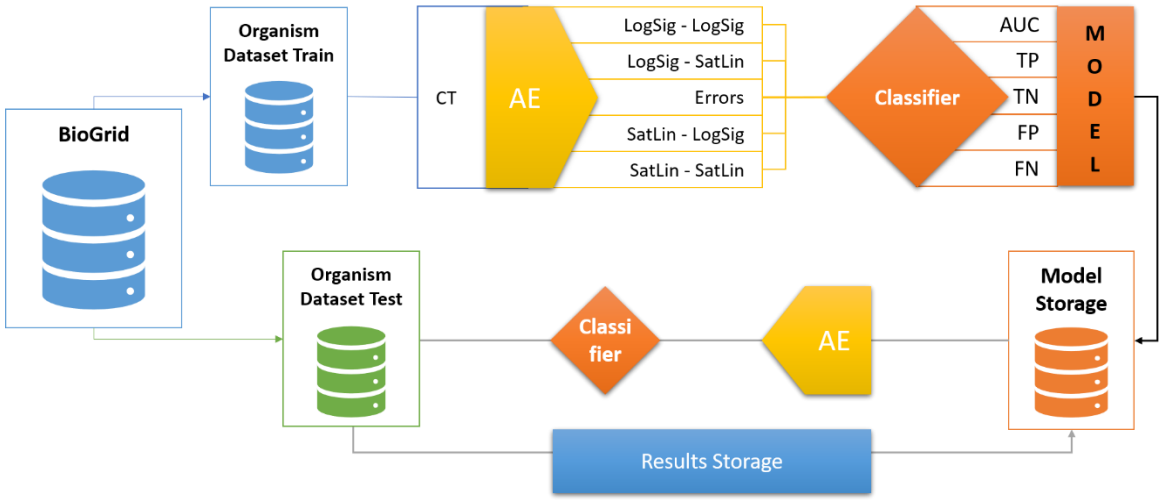

Figure 17 General Scheme of Experiments with BioGrid dataset, using Conjoint Triad and storage of models to be tested in other organisms.

In the second Stage of Experiments, with BioGrid dataset, there are trials regarding the Classifier Training Experiment and the Interspecies & Benchmark Experiment. Although already shown that the Conjoint Triad Method, and the Autoencoders, were chosen as the methods to be applied in the first two experiments, evidences will be presented as to why they were chosen.

The sequence followed in this second stage, as shown by Figure 17 is started by the extraction of an organism as Training Dataset in which 70% of the interactions are selected. The interactions are encoded using Conjoint Triad Method which is delivered to an Autoencoder. This Autoencoder is

trained using all 4 combinations described before, in order to obtain the Hidden Representations as well as the Errors. All these final features are trained in a classifier that ensured that the metric is stored within a structure, as well as the model.
Then, a second organism is chosen as testing. If the second organism is the same as the first, the process is followed using the remaining 30% of the interactions. Otherwise, the process tests the model obtained in the first stage in the whole dataset and stores it in a final "Results Storage" structure.

All the aspects regarding the functions of the Autoencoder, the Classifier Training and the Model Storage and use are adapted, since different organisms require different parameters in the same Neural Networks and Classifiers.
To this purpose, we use the hyperparameter optimization in every first stage training, in order to specialize the Autoencoder in the organism intended.

This means that, though the Encoding Method and Metrics maintain through all thesis, we need to evaluate for each organism tested in the Interspecies and Benchmark Experiment the hyperparameters used, with an optimization process.

5.2 Encoding Method Experiment

Encoding Method Experiment was the first designed. The experiment was only realized using the BioGrid dataset, in which we based our model.
Considering the methods chosen in Chapter 2 following the Amino Acid Principles, we considered three main options:

1. Conjoint Triad Method CT
2. Autocovariance Method AC
3. Mixed Encoding Method AC + CT

The experience involved a simple autoencoder testing, with an analysis of the prediction accuracy that the function returns, and the average of 10 experiments for each of the 4 combinations possible and their standard deviation.

First results were obtained regarding the choice of the Encoding Method Experience. EME is the first round of experiments, and the results in Table 3 were obtained.

Encoding Method	*AUC*
AC	0.63811 ± 0.21655
CT	0.77413 ± 0.22512
AC + CT	0.53380 ± 0.26706

Table 3 Results of the Encoding Method Experience with simple SVM

Despite the Mixed Encoding Method being the concatenation of Autocovariance and Conjoint Triad the junction is an average of the double of the dimension, hardening the composition of a separation hyper plan as the Curse of Dimensionality predicts, especially in the datasets with lower number of protein interaction samples.
Following these results, the Encoding Method used was Conjoint Triad, with a fixed-length of 343 features for each protein, as explained before.

5.3 Dimension Reduction Experiment

Dimension Reduction Experiment, or DRE, was the following stage of experiments to be thought. While EME was performed in isolated form, DRE was tested with the Classifier Training Experiment, since the ensemble of both deep-machine learning approaches could result in different ways in their sum, as they would individually. This forced to a joint evaluation of the algorithms that resulted in no prejudice to each experiment and higher confidence in the values obtained.
Using the Conjoint Triad method as Encoding Method, two main approaches were used: Autoencoders and Long Short-Term Memory.

5.3.1 Autoencoders

As described before, Autoencoders manage to find functions that reproduce the input at the output, finding redundancies in data that can be evaluated as independent, but in which some unknown relations may occur.

Considering the results of the contextualization in the study for autoencoders, the general specificities to the Autoencoder training are:

- Hidden Size: 10% of the input size
    - Since each protein is encoded in a 343-double array, then the hidden layer will have size (343 * 2) * 0.10 ≈ 68

- Sparsity Proportion: 0.01

Based on these general parameters, we evaluated all the others, intending to find the best combination possible of encoding and decoding transfer functions.

5.3.1.1 HitPredict dataset

As presented in Figure 16, the main focus of HitPredict dataset was the choice and validation of the Encoding Method, between the three strategies implemented, and their performance in a chosen classifier. This classifier consisted in a Support Vector Machine of linear kernel. The second stage is the one focused on this section.

- Area Under the Curve = 0.9892 ± 0.0131
- Accuracy = 0.9880 ± 0.0122
- Specificity = 0.9894 ± 0.0134
- Sensitivity = 0.9867 ± 0.0129

5.3.1.2 BioGrid dataset

As mentioned before, and present in Figure 17, BioGrid dataset is also used for the training and testing in the Benchmark Experiments.
The graphic at Figure 18 shows the results for all the combinations tried, including the four two-combinations and the error approaches.

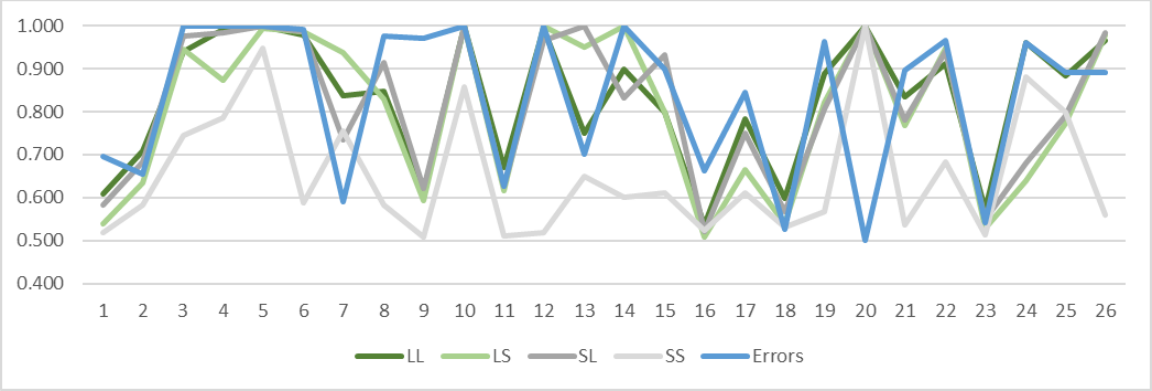

Figure 18 Autoencoders Combinations and Errors AUC

Though these are not the final results of the improved models, there are significant values in some of the organisms analyzed in Figure 18.
In the overall analysis, the results are the ones present on Table 4:

Combination	*LL*	*LS*	*SL*	*SS*	*Errors*
AUC	0.830	0.803	0.816	0.644	0.836
# Best Approach	4	1	3	0	12

Table 4 Mean AUC and best approach of the organisms' results

The final determined approach to the Support Vector Machines Inputs was the Errors. The decisive factors were the Area Under the Curve mean, but also the number of organisms in which the use of errors is the best approach (see Table 5) and the fact that the number of times that the errors are better than the mean of the of the traditional approaches is 22 in the 26 organisms.
A concluding fact is also the dimension inserted in the classifier in which the errors reduce from 68 to only 8 the number of dimensions used.

5.4 Results for the Ensemble Experiment

Despite the selection of a Dimension Reduction technique and a Classifier as two different mechanisms in the same general Pipeline, the results need to be analyzed in the Ensemble all together, to ease the analysis of the implications of the Hidden Representation strategies in the Final Classifiers. This was a necessary effort considering the implications seen in other sections, such as the Encoding Method Experiment, where the method with higher degree of information – Mixed Encoding – produced results below the subsets of Autocovariance and Conjoint Triad.
In these experiments, only the BioGrid dataset was used, and no one configuration for a classifier was used, being described in the different experiments the approach and configurations for each.

5.4.1 Autoencoders + Long Short Term Memory Networks

As a preliminary study, the results found low potential in the development of this structure as a dimension reduction technique, but also as a classifier. In this work, LSTM processed features were the resulting approaches (both error and Hidden Representation) from the Autoencoder. The results obtained, in the accuracy of sequence to label predict, were:

- Using the Hidden Layer from the AE = 0.6105
- Using the 8 Errors Combination from AE = 0.6512

The approach used LSTM was its Sequence to Label form, as the intention was to classify the protein interactions as positive or negative.

5.4.2 Long Short Term Memory + Support Vector Machines

This experiment used LSTM as the dimension reduction technique with LSTM Sequence to Sequence generation. These features were then delivered to the Autoencoder, with implementation of Hyperparameters Optimization in the training process, tuning different parameters of the classier to each different dataset for generalization. The best optimization possible was tried to find, using a Grid Search algorithm in MATLAB Support Vector Machine function to optimize C (Box Constraint) and Gamma (Kernel Scale).

The results obtained were:

- Box Constraint = 1.6481
- Kernel Scale = 261.17
- *Area Under the Curve* = *0.8012*

5.4.3 Autoencoder + Support Vector Machines

The use of autoencoders in the first stage of the ensemble intends to capture the relations between properties of proteins and then feed them to a classifier, in this case Support Vector Machines, that differentiates the overall characteristics of the protein pair. As the two approaches of Hidden Representation and Errors were implemented, a comparison was possible between them, which would validate or undermine the results found before (Table 4) that favored Errors Representation.

Hidden Layer Representations

Using the four representations, an Autoencoder was built with the following result and parameters:

- Box Constraint = 0.0033
- Kernel Scale = Not Aplicable
- Kernel Function = Polinomial, with Order 2
- *Area Under the Curve* = *0.7560*

Errors Representation

As projected in the experiments, the use of the errors of the eight autoencoders built were used in the prediction by SVM, with the following parameters and results:

- Box Constraint = 421.8252
- Kernel Scale = 0.5020
- Kernel Function = Gaussian
- *Area Under the Curve* = *0.9704*

In both approaches, Hyperparameters Optimization was implemented, to which Figure 19 is the representation.

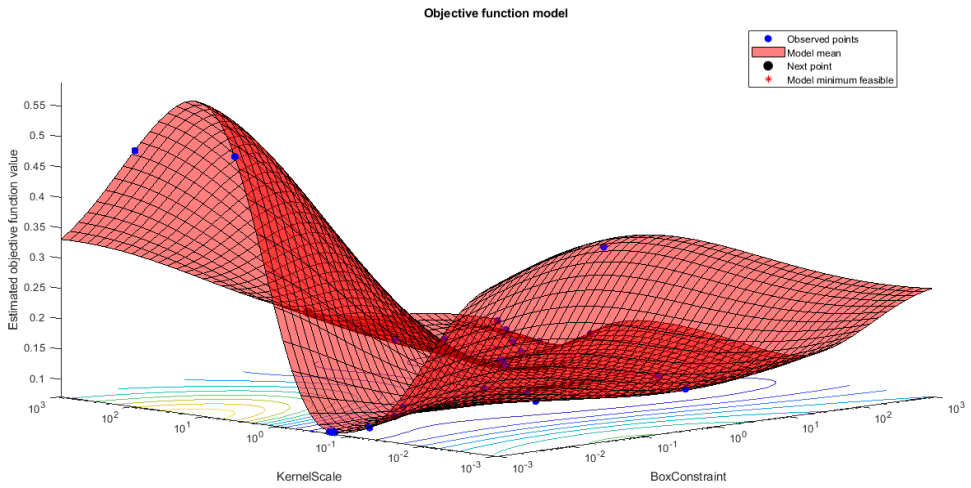

Figure 19 Hyperparameter Optimization in SVM. The Z axis is equal to 1-AUC, while X and Y correspond to the Kernel Scale and Box Constraint tested.

5.4.4 Autoencoder + K-Nearest Neighbors

The principle behind the utilization of autoencoders in the first stage is the same as in the previous ensemble model, and the K-Nearest Neighbors was implemented as a point of reference to the improvement of the prediction of interactions using our approaches (namely Errors and Hidden Representation), since it is one of the already used classifiers in the literature.

Following the same principle of Grid Search in the best approach possible, both Hidden Representation and Errors of Autoencoders as the inputs for the classifier, achieving the results bellow.

Hidden Layer Representations

Using the four representations, we built an Autoencoder with the following results:

- Number of Neighbors 1
- Distance Euclidean
- Distance Weight Equal
- *Area Under the Curve* *0.8534*

Errors Representation

- Number of Neighbors 3
- Distance Euclidean
- Distance Weight Equal
- *Area Under the Curve* *0.9335*

5.4.5 Ensemble general results

Outside of the approach in which Long Short-Term Memory Networks Sequence to Sequence were the first input, all of the ensemble strategies were submitted to both inputs: Errors and Hidden Representation.

The overall results are shown in table 6, where we can distinguish that the highest Area Under the Curve strategy is the Autoencoder + Support vector Machines, with Errors of Representation as features to the classifier.

Ensemble	***AE + LSTM***	***LSTM + SVM***	***AE + SVM***	***AE + KNN***
Hidden Representation	0.610	*0.801*	*0.756*	*0.853*
Errors of Representation	0.651	-----	*0.970*	*0.934*

Table 5 General Table of AUC for the Ensemble

5.5 Classifier Training Experiment

Classifier Training Experiment, or CTE, consisted in the final perfecting of the classifier, in the last stage of the experiments. While nearly all of the solutions performed were tested as described in the previous section, the approach of Long Short Term Memory Sequence to Label, as an overall method that would reduce dimension reduction and classifier in one single step, was not.

5.5.1 Long Short-Term Memory Sequence to Label

As stated before, as Long Short Term Memories have a dedicated sequential approach, their relevance in the field of Bioinformatics is of particular interest in the sequencing experiments. Many encoding methods remove the time sequence feature from the datasets when not present in the original encoding method.

5.5.1.1 Protein related features

In the LSTM approach testing, and using the features described in Chapter 3, we managed to evaluate the contributions that this structure could give to the prediction of interactions.

Predicted	*Real*	
	Interaction	*No Interaction*
Interaction	**2114**	466
No Interaction	386	**2034**
	2500	2500

Table 6 Protein Related features results

Building on the metrics intended to find, as described at Chapter 4, we take the confusion matrix before and calculate the following:

- Accuracy 0.8296
- Specificity 0.8456
- Sensitivity 0.8136

5.5.1.2 Amino acid related features

Using the amino acid related features described in section and using the features as the sequential input to the LSTM Network, we achieved lower results, as shown in confusion matrix below (Table 7).

Predicted	*Real*	
	Interaction	*No Interaction*
Interaction	**1983**	569
No Interaction	517	**1931**
	2500	2500

Table 7 Amino Acid related feature results

Building on the metrics intended to find, as described at Chapter 4, we take the confusion matrix before and calculate the following:

- Accuracy 0.7828
- Specificity 0.7932
- Sensitivity 0.7724

Since the encoding method chosen is of holistic view of the protein, the time feature is already not present. Therefore, to ensure that the characteristics of the amino acids are evaluated in the same way, not losing the time feature, the approach was not implemented, feeding the sequence directly into the LSTM.

Intending to compare with the results obtained in other classifying methods, we built experiments in the HitPredict dataset with 5 000 samples in the testing dataset.

Classifier	*SVM*				*LSTM*
Encoding	CT	AC	CT + AE	AC + AE	----
AUC	0.762	0.711	0.801	0.725	0.775

Table 8 Area Under the Curve using Substitution of Amino Acids vs SVM

5.6 Interspecies and Benchmark Experiment

In the aftermath of the HitPredict Experiments and the first experiments with BioGrid dataset, a hypothesis was theorized about the possibility of predicting Proteins Interactions between organisms, with the goal of predicting which proteins may interact in different organisms, and which proteins, from different organisms, could interact in each organism.

The main challenges in interspecies relations are related to the difference in study and complexity of organisms. The most evidential example is the yeast (Saccharomyces cerevisiae), an organism

less complex than most mammals, but due to its interest in investigation, it has much more data and, specifically, interactions detected and tested.

The research in this area, of interspecies relations, has been the object of different approaches, but is a recent field, with low data comparisons to adequate.

In this stage of experiments, both K-Nearest Neighbors algorithms and Support Vector Machines were used, as a mean of comparison to assess if the capabilities of discrimination have the same degree of improvement in interspecies PPI as they have in intraspecies PPI.
In both experiments, the application of methods was similar. The experiment involved a cycle of testing the models developed in the Classifier Training Experiment in each organism.

5.6.1 K-Nearest Neighbors Classifier

The K-Nearest Neighbors Classifier is one of the most used in the literature, and in the development of the Interspecies Experiment, it was tested and improved using Hyperparameter Optimization.

The first test was made using the five biggest organisms in the dataset. This first sample was of high improvement and Grid Search, achieving the results found in Table 8.

The organisms found in the table, by species ID's, have the following correspondence:

- 3702 Arabidopsis thaliana Arabidopsia
- 7227 Drosophila Monogaster Fruit Fly
- 9606 Homo Sapines Human dataset
- 10090 Mus Muculus Domestic Mouse
- 559292 Sacharomisae Cerevisae Yeast

KNN Experiment

		Test dataset				
		3702	*7227*	*9606*	*10090*	*559292*
Training dataset	*3702*	**0.934**	0.751	0.504	0.671	0.595
	7227	0.543	**0.949**	0.525	0.512	0.980
	9606	0.809	0.575	**0.853**	0.722	0.875
	10090	0.568	0.558	0.501	**0.849**	0.838
	559292	0.875	0.596	0.543	0.878	**0.851**

Table 9 Interspecies AUC of the 5 main organisms' datasets (K-NN)

As the results allowed for low extrapolations while attaining high accuracies in some of the cases, we augmented the sample to all the approved 26 organisms' datasets, to which we achieved the results, in color table, in Table 10. The table with higher number of decimal precision points is present in Annex 6.

	10090	10116	10238	10359	10376	11676	227321	237561	284812	333760	3702	37296	3847	39947	511145	559292	6239	7227	7955	83332	8355	9031	9606	9823	9913	9986
10090	0.849	0.529	0.693	0.807	0.897	0.660	0.609	0.901	0.678	0.907	0.568	0.743	0.757	0.780	0.893	0.838	0.808	0.558	0.796	0.516	0.702	0.879	0.501	0.728	0.635	0.525
10116	0.594	0.677	0.523	0.841	0.863	0.655	0.683	0.578	0.510	0.904	0.507	0.744	0.914	0.913	0.746	0.639	0.621	0.553	0.706	0.649	0.810	0.502	0.614	0.761	0.503	0.675
10298	0.503	0.535	0.828	0.864	0.904	0.774	0.500	0.506	0.641	0.811	0.593	0.972	0.607	0.788	0.716	0.643	0.530	0.658	0.867	0.564	0.848	0.752	0.504	0.785	0.581	0.583
10359	0.566	0.541	0.837	0.953	0.982	0.790	0.501	0.612	0.501	0.818	0.570	0.978	0.753	0.792	0.503	0.639	0.592	0.509	0.879	0.523	0.780	0.562	0.662	0.502	0.577	0.797
10376	0.570	0.587	0.960	0.800	0.985	0.664	0.621	0.587	0.535	0.928	0.592	0.929	0.696	0.814	0.746	0.546	0.524	0.573	0.851	0.542	0.680	0.785	0.510	0.708	0.605	0.792
11676	0.536	0.576	0.785	0.849	0.911	0.888	0.783	0.796	0.552	0.891	0.609	0.626	0.912	0.641	0.765	0.594	0.528	0.650	0.576	0.510	0.512	0.811	0.570	0.726	0.538	0.812
227321	0.659	0.583	0.551	0.523	0.691	0.510	0.608	0.695	0.557	0.629	0.596	0.511	0.529	0.740	0.766	0.583	0.547	0.539	0.589	0.571	0.609	0.640	0.527	0.687	0.569	0.715
237561	0.563	0.516	0.584	0.713	0.612	0.660	0.622	0.825	0.628	0.699	0.505	0.818	0.714	0.739	0.739	0.580	0.618	0.681	0.559	0.597	0.580	0.827	0.597	0.881	0.503	0.732
284812	0.533	0.570	0.564	0.528	0.778	0.541	0.612	0.694	0.852	0.753	0.613	0.526	0.616	0.570	0.553	0.661	0.525	0.622	0.544	0.539	0.523	0.592	0.525	0.668	0.616	0.622
333760	0.521	0.539	0.787	0.874	0.914	0.528	0.576	0.800	0.521	0.922	0.562	0.798	0.677	0.891	0.561	0.585	0.545	0.539	0.694	0.531	0.540	0.653	0.532	0.560	0.551	0.705
3702	0.671	0.574	0.654	0.633	0.609	0.526	0.569	0.634	0.608	0.899	0.934	0.810	0.689	0.875	0.715	0.595	0.687	0.751	0.585	0.586	0.553	0.739	0.504	0.885	0.579	0.682
37296	0.608	0.514	0.843	0.898	0.808	0.709	0.512	0.694	0.543	0.838	0.577	0.789	0.779	0.865	0.687	0.589	0.774	0.654	0.748	0.526	0.695	0.596	0.510	0.503	0.522	0.694
3847	0.608	0.648	0.976	0.893	0.956	0.743	0.671	0.838	0.613	0.985	0.546	0.945	0.684	0.934	0.779	0.650	0.582	0.672	0.876	0.679	0.812	0.879	0.571	0.779	0.734	0.574
39947	0.670	0.602	0.789	0.886	0.925	0.544	0.619	0.644	0.553	0.908	0.667	0.784	0.633	0.969	0.581	0.600	0.578	0.506	0.801	0.618	0.676	0.682	0.642	0.536	0.588	0.789
511145	0.561	0.514	0.836	0.949	0.808	0.775	0.669	0.755	0.724	0.836	0.605	0.990	0.680	0.705	0.638	0.626	0.651	0.656	0.846	0.617	0.746	0.807	0.630	0.797	0.735	0.598
559292	0.878	0.536	0.729	0.691	0.512	0.526	0.544	0.538	0.641	0.669	0.875	0.745	0.517	0.502	0.607	0.851	0.547	0.597	0.658	0.566	0.648	0.615	0.543	0.525	0.508	0.652
6239	0.585	0.536	0.882	0.861	0.977	0.591	0.588	0.563	0.609	0.814	0.511	0.711	0.794	0.830	0.538	0.588	0.576	0.501	0.641	0.576	0.771	0.626	0.503	0.561	0.561	0.528
7227	0.512	0.527	0.574	0.832	0.706	0.594	0.555	0.678	0.697	0.939	0.543	0.589	0.636	0.662	0.528	0.981	0.584	0.949	0.665	0.596	0.560	0.505	0.525	0.591	0.626	0.819
7955	0.581	0.526	0.643	0.850	0.794	0.616	0.575	0.543	0.527	0.814	0.598	0.671	0.733	0.924	0.660	0.525	0.552	0.554	0.676	0.596	0.569	0.626	0.648	0.517	0.508	0.604
83332	0.540	0.510	0.821	0.762	0.925	0.503	0.581	0.697	0.693	0.882	0.607	0.779	0.642	0.894	0.555	0.571	0.563	0.508	0.731	0.642	0.616	0.604	0.586	0.562	0.519	0.796
8355	0.562	0.506	0.876	0.803	0.830	0.762	0.536	0.639	0.650	0.881	0.601	0.966	0.818	0.842	0.773	0.609	0.638	0.616	0.866	0.588	0.773	0.748	0.589	0.775	0.729	0.745
9031	0.613	0.562	0.859	0.797	0.814	0.617	0.574	0.893	0.778	0.816	0.634	0.943	0.763	0.748	0.865	0.500	0.662	0.523	0.814	0.546	0.722	0.715	0.635	0.751	0.716	0.573
9606	0.722	0.525	0.673	0.786	0.879	0.665	0.572	0.616	0.698	0.615	0.809	0.532	0.628	0.679	0.757	0.875	0.540	0.575	0.643	0.687	0.692	0.700	0.853	0.747	0.532	0.619
9823	0.579	0.531	0.897	0.819	0.975	0.896	0.657	0.643	0.546	0.848	0.581	0.934	0.710	0.800	0.683	0.582	0.653	0.692	0.672	0.538	0.787	0.722	0.687	0.698	0.640	0.545
9913	0.515	0.590	0.807	0.826	0.958	0.783	0.503	0.787	0.505	0.849	0.516	0.830	0.646	0.751	0.805	0.532	0.608	0.517	0.638	0.590	0.692	0.705	0.663	0.755	0.558	0.556
9986	0.531	0.538	0.518	0.776	0.713	0.807	0.660	0.586	0.520	0.980	0.654	0.695	0.837	0.606	0.602	0.565	0.628	0.500	0.631	0.635	0.528	0.555	0.505	0.598	0.618	0.777

Table 10 AUC of Interspecies testing with K-NN

5.6.2 Support Vector Machines

With basis in this table, the experiments described in the Interspecies & Benchmark Experiment were designed to correspond to the same organisms. Considering this effort, the following results were obtained from Interspecies Experiment.

Since the experiment conducted attained good results, it was extended to the other 26 datasets with the results, in color table, in Table 12. The table with higher number of decimal precision points is present in Annex 7.

Interspecies Trial

		Test dataset				
		3702	*7227*	*9606*	*10090*	*559292*
Training dataset	*3702*	**0.996**	0.811	0.599	0.825	0.658
	7227	0.643	**0.973**	0.672	0.544	0.989
	9606	0.983	0.516	**0.970**	0.870	0.957
	10090	0.756	0.612	0.514	**0.863**	0.893
	559292	0.932	0.782	0.639	0.970	**0.964**

Table 11 Interspecies AUC of the 5 main organisms' datasets (SVM)

To prove the model in all its approaches, another round of verifications was made, without Hyperparameter Optimization, in which all the approaches regarding the combination of Encoder and Decoding Transfer Functions were tested and assessed in AUC.

	10090	10116	10238	10359	10376	11676	227321	237561	284812	333760	3702	37296	3847	39947	511145	559292	6239	7227	7955	83332	8355	9031	9606	9823	9913	9986
10090	0.863	0.599	0.728	0.993	0.934	0.856	0.773	0.902	0.861	1.000	0.756	0.884	0.792	0.852	0.948	0.893	0.820	0.612	0.859	0.563	0.820	0.947	0.514	0.902	0.771	0.711
10116	0.633	0.802	0.652	0.849	0.922	0.841	0.717	0.533	0.524	0.989	0.554	0.837	0.917	0.914	0.812	0.512	0.796	0.535	0.778	0.500	0.825	0.654	0.505	0.884	0.631	0.689
10298	0.578	0.588	0.984	1.000	0.997	0.895	0.694	0.693	0.767	1.000	0.547	0.998	0.750	0.957	0.803	0.510	0.708	0.518	0.944	0.625	0.865	0.904	0.514	0.898	0.702	0.739
10359	0.592	0.592	0.990	0.981	0.985	0.853	0.649	0.639	0.541	1.000	0.600	0.999	0.875	0.965	0.621	0.553	0.764	0.509	0.900	0.625	0.810	0.725	0.508	0.578	0.642	0.870
10376	0.572	0.599	0.981	0.999	0.995	0.841	0.680	0.736	0.588	1.000	0.554	1.000	0.854	0.961	0.849	0.507	0.723	0.514	0.894	0.500	0.860	0.865	0.501	0.836	0.747	0.797
11676	0.601	0.526	0.979	0.985	0.970	0.919	0.902	0.805	0.506	1.000	0.519	0.665	0.979	0.543	0.796	0.552	0.653	0.511	0.588	0.500	0.513	0.933	0.515	0.911	0.546	0.896
227321	0.501	0.533	0.605	0.602	0.805	0.662	0.643	0.887	0.721	0.511	0.543	0.678	0.583	0.770	0.794	0.528	0.563	0.501	0.519	0.563	0.564	0.756	0.501	0.853	0.573	0.791
237561	0.609	0.619	0.717	0.746	0.728	0.530	0.528	0.944	0.821	0.889	0.537	0.936	0.743	0.758	0.776	0.515	0.690	0.509	0.617	0.563	0.705	0.939	0.522	0.938	0.639	0.856
284812	0.603	0.575	0.626	0.581	0.830	0.567	0.540	0.753	0.876	0.885	0.528	0.661	0.771	0.531	0.591	0.512	0.564	0.509	0.726	0.563	0.696	0.763	0.527	0.849	0.523	0.669
333760	0.532	0.570	0.928	0.975	0.967	0.652	0.569	0.812	0.647	1.000	0.504	0.865	0.715	0.934	0.580	0.589	0.517	0.532	0.860	0.563	0.539	0.660	0.516	0.760	0.513	0.883
3702	0.825	0.551	0.504	0.732	0.515	0.657	0.576	0.671	0.550	0.910	0.996	0.863	0.806	0.918	0.866	0.658	0.688	0.811	0.609	0.563	0.559	0.861	0.599	0.920	0.556	0.726
37296	0.616	0.614	0.950	0.992	0.982	0.796	0.680	0.502	0.570	1.000	0.569	0.983	0.847	0.871	0.706	0.541	0.779	0.520	0.863	0.563	0.785	0.781	0.520	0.600	0.671	0.843
3847	0.572	0.672	0.998	1.000	0.997	0.823	0.503	0.866	0.532	1.000	0.584	0.972	0.800	0.977	0.813	0.507	0.708	0.517	0.887	0.719	0.830	0.882	0.508	0.969	0.778	0.657
39947	0.512	0.512	0.958	0.997	0.997	0.728	0.655	0.531	0.574	1.000	0.532	0.978	0.757	1.000	0.770	0.544	0.593	0.527	0.905	0.750	0.778	0.751	0.511	0.542	0.670	0.819
511145	0.642	0.609	0.986	0.999	0.978	0.906	0.739	0.932	0.811	1.000	0.583	0.995	0.840	0.895	0.733	0.553	0.779	0.507	0.869	0.563	0.831	0.960	0.521	0.960	0.817	0.599
559292	0.970	0.515	0.771	0.738	0.650	0.646	0.553	0.649	0.529	0.858	0.932	0.832	0.500	0.570	0.530	0.964	0.643	0.782	0.806	0.500	0.547	0.552	0.639	0.556	0.581	0.685
6239	0.646	0.550	0.980	0.955	0.994	0.752	0.698	0.640	0.563	0.994	0.589	0.892	0.903	0.832	0.716	0.550	0.706	0.530	0.834	0.625	0.841	0.719	0.506	0.604	0.634	0.694
7227	0.544	0.516	0.582	0.947	0.866	0.663	0.583	0.689	0.768	1.000	0.643	0.589	0.701	0.828	0.688	0.989	0.531	0.973	0.728	0.500	0.569	0.641	0.672	0.680	0.523	0.838
7955	0.535	0.514	0.780	0.938	0.956	0.558	0.549	0.531	0.618	0.956	0.539	0.844	0.799	0.973	0.667	0.645	0.533	0.536	0.855	0.625	0.680	0.666	0.521	0.578	0.569	0.594
83332	0.545	0.554	0.885	0.957	0.964	0.613	0.608	0.727	0.744	1.000	0.528	0.799	0.708	0.918	0.620	0.617	0.533	0.543	0.756	0.500	0.510	0.588	0.500	0.751	0.503	0.871
8355	0.564	0.639	0.989	1.000	0.997	0.829	0.630	0.731	0.548	1.000	0.573	0.999	0.840	1.000	0.857	0.500	0.716	0.521	0.918	0.563	0.800	0.908	0.512	0.889	0.772	0.759
9031	0.637	0.600	0.989	0.977	0.977	0.801	0.662	0.933	0.866	0.919	0.560	0.995	0.792	0.883	0.958	0.521	0.762	0.528	0.866	0.563	0.817	0.867	0.519	0.938	0.789	0.553
9606	0.870	0.527	0.767	0.924	0.912	0.512	0.522	0.662	0.847	0.801	0.983	0.596	0.778	0.852	0.772	0.957	0.538	0.516	0.813	0.813	0.730	0.715	0.970	0.849	0.604	0.652
9823	0.632	0.613	0.985	0.994	1.000	0.916	0.662	0.821	0.684	1.000	0.579	1.000	0.875	0.875	0.839	0.549	0.809	0.500	0.852	0.500	0.817	0.859	0.508	0.800	0.783	0.699
9913	0.569	0.600	0.985	1.000	0.979	0.833	0.676	0.873	0.517	1.000	0.587	0.989	0.750	0.898	0.961	0.597	0.699	0.513	0.763	0.500	0.802	0.892	0.504	0.942	0.753	0.633
9986	0.508	0.569	0.521	0.960	0.890	0.815	0.719	0.707	0.622	1.000	0.526	0.877	0.910	0.715	0.704	0.535	0.675	0.534	0.718	0.563	0.673	0.747	0.502	0.516	0.642	0.895

Table 12 AUC of Interspecies testing with SVM

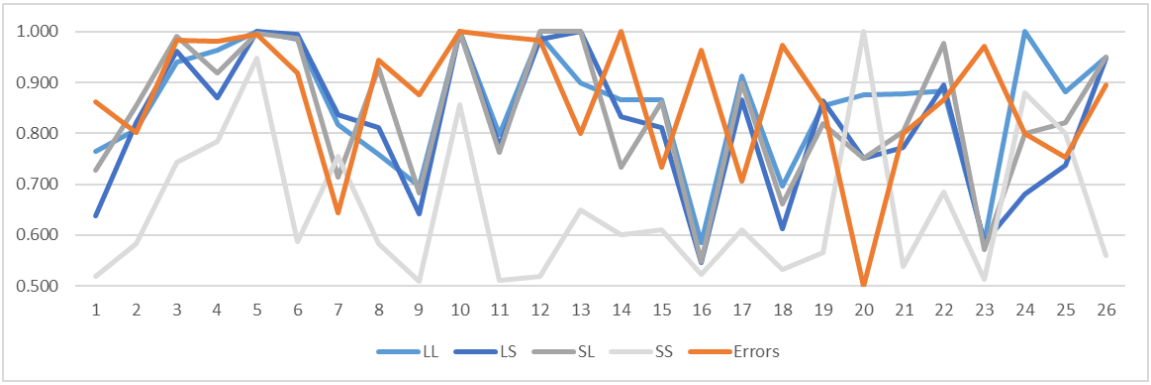

Figure 20 Validation Results for the Combinations and Errors approach

We can therefore conclude that, for the mean of all datasets tested, and the four combinations and errors approach, the one with the highest mean AUC is the Errors, as validated by the last results in Figure 20.

Combination	*LL*	*LS*	*SL*	*SS*	*Errors*
AUC	0,856	0,817	0,837	0,644	0,858

Table 13 Mean results of AUC for all organisms

6 Discussion and Conclusions

After the results obtained from the experiments in previous chapter, an overview of the strategies implemented, and their extent is necessary to conclude the best possible model for the prediction of Interspecies PPI.
This chapter focus on those results comparing them to the prevailing literature and State of the Art, and addressing some data retrieved but not analyzed in the previous chapters, coordinating them with both biological aspects, and the objectives of this thesis.

6.1 Protein-Protein Interaction Prediction

The first objective was the production of a model that could predict Protein Interactions in the Human Organism. This goal was achieved based on model-based predictions using the HitPredict dataset and testing them in BioGRID as a benchmark dataset.

The novelty in our approach is the use of Autoencoder Errors to predict these interactions, instead of using the traditional Hidden Representation approach. Comparing both approaches, all results lead to the conclusion that the use of errors attains more information. In the eyes of a computer scientist, this means a flip of tables in the way an autoencoder can be used for classification and not as a reduced representation. This implies that a given point in high dimensional space, if evaluated from different perspectives, can be viewed as a class belonging point or not, in which the Autoencoder is the center of the class or cluster.

If the point is near the cluster (low error) it means that the Autoencoder (positive or negative trained) predicts with high precision that point, therefore it may belong to its class (positive or negative). If, however, the point has an high value error, it may imply that it does not belong to the class in which the Autoencoder was trained.

This simple view must be then added three times, since this approach is used in the four possible combinations described in 4.5 Autoencoder Approach. As this approach is used only 2 times in 4 different perspectives, the number of features that the classifier needs to differentiate is low, giving advantage in the hyper plan determination when compared to the 68-feature input space that the Hidden Representation strategy has.

The results obtained are not the best intended for our datasets, and represent preliminary hypothesis, to be tested with further efforts. However, as Table 14 shows, they are comparable to the State of the Art, surpassed only by the Stacked Autoencoder method by Sun et al. 2017 [52], some information can still be retrieved from them and how to proceed with the models developed.

Reference	*Algorithm*	*Training ACC*
J. Shen et al [48]	SVM	0.830
Y. Guo et al [53]	SVM	0.904
X.-Y. Pan [54]	LDA-RDF	0.979
Y.-N. Zhang [55]	CS-SVM	0.940
Z. You [56]	ELM	0.848
Z. You [57]	SVM	0.920 – 0.974
T. Sun [52]	SAE	0.972
Our Model	AE (errors) - SVM	0.989

Table 14 Accuracies of the work developed in HitPredict dataset by comparison with other state of the art methods

The Pipeline designed takes shape by forming a connection between organisms. And the connection is obvious between some characteristics of the datasets found.

6.2 Organism's Specificities

During the experiments several data were stored in order to analyze them in the Conclusions, not only related to a single organism, but to the relations between them.

As we can see in Figure 21, there is a mixed relation between the sizes of the datasets and their AUC. This data is coherent with the fact that smaller datasets are easier to fit into a classification model, but bigger datasets also have higher discrimination power.

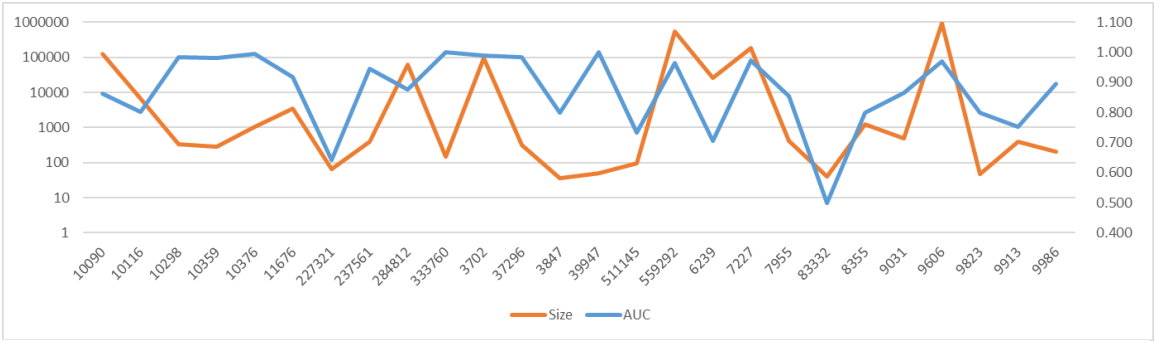

Figure 21 Graphic of the Size of the dataset (base 10 axis) and respective AUC by organism

The disparity lead to the search for the phylogenetic relationships. Phylogeny is one of the measures of similarity between organisms, being related to common ancestors, and to which protein interactions and common proteins is one of the measurable features.

Considering this, we developed a Phylogenetic tree, resorting to the web platform TimeTree [58], in which the input species names generate a phylogeny tree, shown in Figure 22.

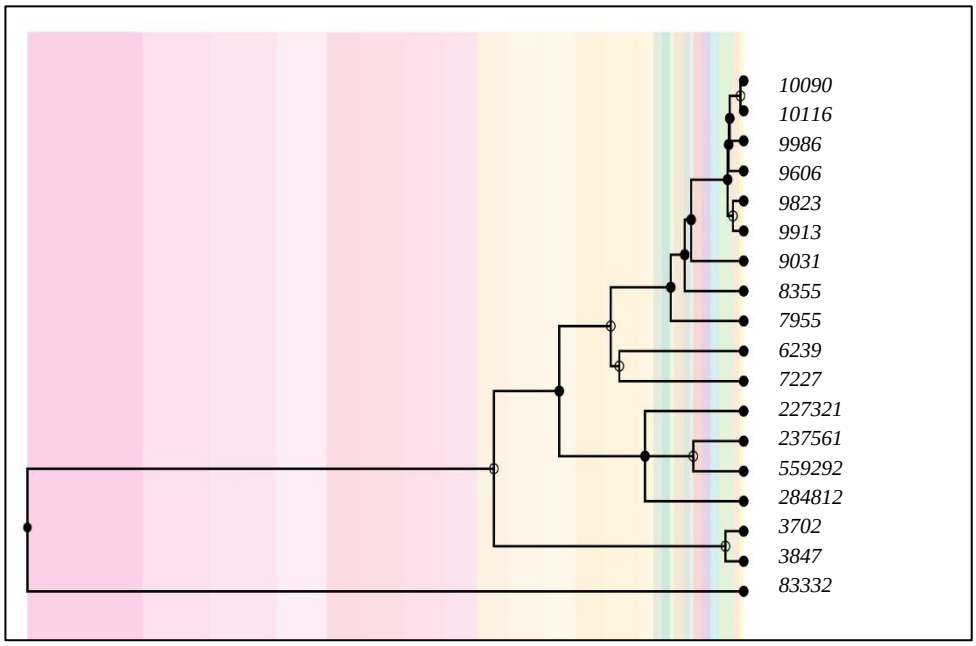

Figure 22 Phylogenetic relationship between organisms considered resorting to the TimeTree web platform.

The intention was to find if these relationships could explain some of the disparity in some of the organisms. Using the known organisms and their phylogenetic relationship, we rearranged the results obtained in the Interspecies and Benchmark Experiment, obtaining the following Table 15 for the SVM classifier.

	10090	10116	9986	9606	9823	9913	9031	8355	7955	6239	7227	227321	237561	559292	284812	3702	3847	83332
10090	0.863	0.599	0.711	0.514	0.902	0.771	0.947	0.820	0.859	0.820	0.612	0.773	0.902	0.893	0.861	0.756	0.792	0.563
10116	0.633	0.802	0.689	0.505	0.884	0.631	0.654	0.825	0.778	0.796	0.535	0.717	0.533	0.512	0.524	0.554	0.917	0.500
9986	0.508	0.569	0.895	0.502	0.516	0.642	0.747	0.673	0.718	0.675	0.534	0.719	0.707	0.535	0.622	0.526	0.910	0.563
9606	0.870	0.527	0.652	0.970	0.849	0.604	0.715	0.730	0.813	0.538	0.516	0.522	0.662	0.957	0.847	0.983	0.778	0.813
9823	0.632	0.613	0.699	0.508	0.800	0.783	0.859	0.817	0.852	0.809	0.500	0.662	0.821	0.549	0.684	0.579	0.875	0.500
9913	0.569	0.600	0.633	0.504	0.942	0.753	0.892	0.802	0.763	0.699	0.513	0.676	0.873	0.597	0.517	0.587	0.750	0.500
9031	0.637	0.600	0.553	0.519	0.938	0.789	0.867	0.817	0.866	0.762	0.528	0.662	0.933	0.521	0.866	0.560	0.792	0.563
8355	0.564	0.639	0.759	0.512	0.889	0.772	0.908	0.800	0.918	0.716	0.521	0.630	0.731	0.500	0.548	0.573	0.840	0.563
7955	0.535	0.514	0.594	0.521	0.578	0.569	0.666	0.680	0.855	0.533	0.536	0.549	0.531	0.645	0.618	0.539	0.799	0.625
6239	0.646	0.550	0.694	0.506	0.604	0.634	0.719	0.841	0.834	0.706	0.530	0.698	0.640	0.550	0.563	0.589	0.903	0.625
7227	0.544	0.516	0.838	0.672	0.680	0.523	0.641	0.569	0.728	0.531	0.973	0.583	0.689	0.989	0.768	0.643	0.701	0.500
227321	0.501	0.533	0.791	0.501	0.853	0.573	0.756	0.564	0.519	0.563	0.501	0.643	0.887	0.528	0.721	0.543	0.583	0.563
237561	0.609	0.619	0.856	0.522	0.938	0.639	0.939	0.705	0.617	0.690	0.509	0.528	0.944	0.515	0.821	0.537	0.743	0.563
559292	0.970	0.515	0.685	0.639	0.556	0.581	0.552	0.547	0.806	0.643	0.782	0.553	0.649	0.964	0.529	0.932	0.500	0.500
284812	0.603	0.575	0.669	0.527	0.849	0.523	0.763	0.696	0.726	0.564	0.509	0.540	0.753	0.512	0.876	0.528	0.771	0.563
3702	0.825	0.551	0.726	0.599	0.920	0.556	0.861	0.559	0.609	0.688	0.811	0.576	0.671	0.658	0.550	0.996	0.806	0.563
3847	0.572	0.672	0.657	0.508	0.969	0.778	0.882	0.830	0.887	0.708	0.517	0.503	0.866	0.507	0.532	0.584	0.800	0.719
83332	0.545	0.554	0.871	0.500	0.751	0.503	0.588	0.510	0.756	0.533	0.543	0.608	0.727	0.617	0.744	0.528	0.708	0.500

Table 15 AUC of Interspecies testing with SVM ordered by phylogenetic relationship.

The expected result would be a confluence of higher AUC in the diagonal of the Table, which is not the obtained. We know, from previous results in the experiments that the relationship between datasets follows the relation in two main points:

1. Organisms with larger dataset size predict small and medium size dataset organisms, as well as themselves, with high accuracy, but not large size dataset organisms;
2. Organisms with lower dataset size cannot predict medium or large size organism, but can predict organisms of similar structure and themselves;

But the quantification of these relationships could bring some explanation to the obtained results. In the case of the size of the datasets, the relation between the datasets was evaluated by the logarithm of the proportion between datasets (Table 16) and the Phylogenetic distance was evaluated using the binary encoding of the nodes, in which each node represents a digit 1 or 0 considering it moves down or up from the node it originates (Table 17). Therefore, the tables obtained are presented below.

	10090	10116	9986	9606	9823	9913	9031	8355	7955	6239	7227	227321	237561	559292	284812	3702	3847	83332
10090	0.000	1.509	4.996	-0.319	6.480	4.335	4.119	3.209	4.262	0.125	0.000	6.135	4.338	0.000	0.000	0.000	6.725	6.595
10116	-1.509	0.000	3.487	-1.828	4.972	2.827	2.610	1.700	2.753	-1.384	-1.509	4.626	2.829	-1.509	-1.509	-1.509	5.217	5.087
9986	-4.996	-3.487	0.000	-5.315	1.485	-0.661	-0.877	-1.787	-0.734	-4.871	-4.996	1.139	-0.658	-4.996	-4.996	-4.996	1.730	1.600
9606	0.319	1.828	5.315	0.000	6.799	4.654	4.438	3.528	4.581	0.444	0.319	6.454	4.657	0.319	0.319	0.319	7.044	6.914
9823	-6.480	-4.972	-1.485	-6.799	0.000	-2.145	-2.362	-3.271	-2.219	-6.356	-6.480	-0.346	-2.143	-6.480	-6.480	-6.480	0.245	0.115
9913	-4.335	-2.827	0.661	-4.654	2.145	0.000	-0.217	-1.126	-0.074	-4.210	-4.335	1.799	0.003	-4.335	-4.335	-4.335	2.390	2.260
9031	-4.119	-2.610	0.877	-4.438	2.362	0.217	0.000	-0.910	0.143	-3.994	-4.119	2.016	0.219	-4.119	-4.119	-4.119	2.607	2.477
8355	-3.209	-1.700	1.787	-3.528	3.271	1.126	0.910	0.000	1.053	-3.084	-3.209	2.926	1.129	-3.209	-3.209	-3.209	3.517	3.386
7955	-4.262	-2.753	0.734	-4.581	2.219	0.074	-0.143	-1.053	0.000	-4.137	-4.262	1.873	0.076	-4.262	-4.262	-4.262	2.464	2.334
6239	-0.125	1.384	4.871	-0.444	6.356	4.210	3.994	3.084	4.137	0.000	-0.125	6.010	4.213	-0.125	-0.125	-0.125	6.601	6.471
7227	0.000	1.509	4.996	-0.319	6.480	4.335	4.119	3.209	4.262	0.125	0.000	6.135	4.338	0.000	0.000	0.000	6.725	6.595
227321	-6.135	-4.626	-1.139	-6.454	0.346	-1.799	-2.016	-2.926	-1.873	-6.010	-6.135	0.000	-1.797	-6.135	-6.135	-6.135	0.591	0.461
237561	-4.338	-2.829	0.658	-4.657	2.143	-0.003	-0.219	-1.129	-0.076	-4.213	-4.338	1.797	0.000	-4.338	-4.338	-4.338	2.388	2.258
559292	0.000	1.509	4.996	-0.319	6.480	4.335	4.119	3.209	4.262	0.125	0.000	6.135	4.338	0.000	0.000	0.000	6.725	6.595
284812	0.000	1.509	4.996	-0.319	6.480	4.335	4.119	3.209	4.262	0.125	0.000	6.135	4.338	0.000	0.000	0.000	6.725	6.595
3702	0.000	1.509	4.996	-0.319	6.480	4.335	4.119	3.209	4.262	0.125	0.000	6.135	4.338	0.000	0.000	0.000	6.725	6.595
3847	-6.725	-5.217	-1.730	-7.044	-0.245	-2.390	-2.607	-3.517	-2.464	-6.601	-6.725	-0.591	-2.388	-6.725	-6.725	-6.725	0.000	-0.130
83332	-6.595	-5.087	-1.600	-6.914	-0.115	-2.260	-2.477	-3.386	-2.334	-6.471	-6.595	-0.461	-2.258	-6.595	-6.595	-6.595	0.130	0.000

Table 16 Natural logarithm of the proportion between the datasets' size.

	1022	10116	9986	9606	9823	9913	9031	8355	7955	6239	7227	227321	237561	559292	284812	3702	3847	83332
10090	0.000	0.000	0.695	1.390	2.087	2.788	2.788	3.497	4.223	4.257	4.292	5.059	4.328	4.366	5.139	6.015	6.238	7.624
10116	0.000	0.000	0.694	1.389	2.086	2.787	2.787	3.497	4.222	4.256	4.291	5.059	4.328	4.366	5.139	6.014	6.237	7.624
9986	-0.695	-0.694	0.000	0.695	1.392	2.093	2.093	2.802	3.528	3.562	3.597	4.365	3.634	3.671	4.445	5.320	5.543	6.930
9606	-1.390	-1.389	-0.695	0.000	0.697	1.398	1.398	2.107	2.833	2.867	2.902	3.669	2.939	2.976	3.750	4.625	4.848	6.234
9823	-2.087	-2.086	-1.392	-0.697	0.000	0.701	0.701	1.410	2.136	2.170	2.205	2.972	2.241	2.279	3.052	3.928	4.151	5.537
9913	-2.788	-2.787	-2.093	-1.398	-0.701	0.000	0.000	0.709	1.435	1.469	1.504	2.271	1.540	1.578	2.351	3.227	3.450	4.836
9031	-2.788	-2.787	-2.093	-1.398	-0.701	0.000	0.000	0.709	1.435	1.469	1.504	2.271	1.540	1.578	2.351	3.227	3.450	4.836
8355	-3.497	-3.497	-2.802	-2.107	-1.410	-0.709	-0.709	0.000	0.726	0.760	0.795	1.562	0.831	0.869	1.642	2.518	2.741	4.127
7955	-4.223	-4.222	-3.528	-2.833	-2.136	-1.435	-1.435	-0.726	0.000	0.034	0.069	0.836	0.105	0.143	0.916	1.792	2.015	3.401
6239	-4.257	-4.256	-3.562	-2.867	-2.170	-1.469	-1.469	-0.760	-0.034	0.000	0.035	0.802	0.071	0.109	0.882	1.758	1.981	3.367
7227	-4.292	-4.291	-3.597	-2.902	-2.205	-1.504	-1.504	-0.795	-0.069	-0.035	0.000	0.767	0.036	0.074	0.847	1.723	1.946	3.332
227321	-5.059	-5.059	-4.365	-3.669	-2.972	-2.271	-2.271	-1.562	-0.836	-0.802	-0.767	0.000	-0.731	-0.693	0.080	0.956	1.179	2.565
237561	-4.328	-4.328	-3.634	-2.939	-2.241	-1.540	-1.540	-0.831	-0.105	-0.071	-0.036	0.731	0.000	0.038	0.811	1.686	1.910	3.296
559292	-4.366	-4.366	-3.671	-2.976	-2.279	-1.578	-1.578	-0.869	-0.143	-0.109	-0.074	0.693	-0.038	0.000	0.773	1.649	1.872	3.258
284812	-5.139	-5.139	-4.445	-3.750	-3.052	-2.351	-2.351	-1.642	-0.916	-0.882	-0.847	-0.080	-0.811	-0.773	0.000	0.875	1.099	2.485
3702	-6.015	-6.014	-5.320	-4.625	-3.928	-3.227	-3.227	-2.518	-1.792	-1.758	-1.723	-0.956	-1.686	-1.649	-0.875	0.000	0.223	1.609
3847	-6.238	-6.237	-5.543	-4.848	-4.151	-3.450	-3.450	-2.741	-2.015	-1.981	-1.946	-1.179	-1.910	-1.872	-1.099	-0.223	0.000	1.386
83332	-7.624	-7.624	-6.930	-6.234	-5.537	-4.836	-4.836	-4.127	-3.401	-3.367	-3.332	-2.565	-3.296	-3.258	-2.485	-1.609	-1.386	0.000

Table 17 Natural logarithm of the distance between datasets.

In both cases, there is no direct correlation between either one of the tables and Table 15. Therefore, no conclusion can be presented to describe the relationship that these organisms have between them when predicting interactions in interspecies context.

There is, however, some evidence that phylogenetic relationships influence the results, as it is shown in maximum level in the case of *Mycobacterium tuberculosis*, with UniProt ID 83332, which represents itself as a low volume dataset organism (41 positive interactions), but is one of the lowest accurately predicted species in the experiments, inferring the phylogenetic distance that Figure 22 shows. This can be a starting point for further work using these results and the approach developed in this thesis.

6.3 Model Constructed

The final model constructed is the followed Pipeline, to which the best results were obtained in the training and testing datasets. Interspecies protein-protein interaction results are extremely hard to compare, since they are a recent field of work in Bioinformatics.

Taking the best option in each step of the Pipeline, we can create a chronology of the methods to use having as initial step a dataset of protein ID's:

1) Extraction of the sequence and application of the Conjoint Triad Encoding Method;
2) Application of the four autoencoders structure in the post-encoding dataset;
3) Retrieval of the errors, concatenation and training of the SVM;
4) Storage of the model and organism trained and its accuracy.

If intended another approach, and the starting point is an already trained model and a protein pair or a set of protein ID's, the sequence continues:

5) Insertion of a protein ID's pair, and the intended organism;
6) Retrieval of the sequence and preprocessing using Conjoint Triad Method;
7) Using of the stored organism Model to predict the interaction.

Therefore, the built Pipeline related to HitPredict dataset could be synthetized in the flow of Figure 23.

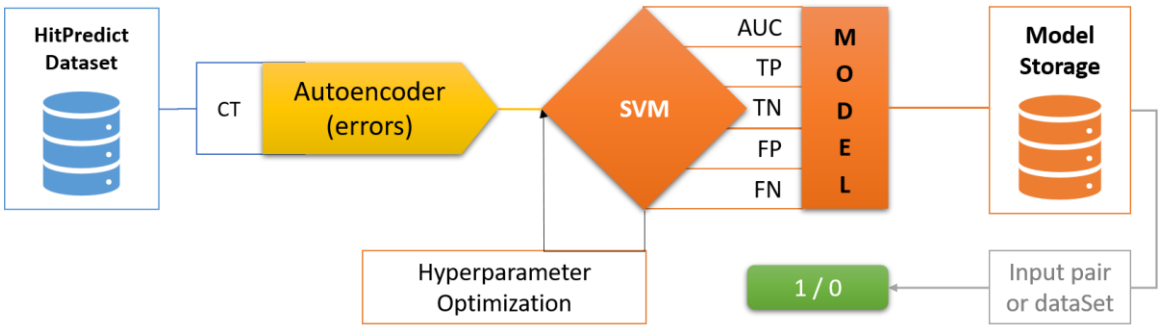

Figure 23 Pipeline generated for one species classification using the HitPredict Dataset, the Conjoint Triad Method and Hyperparameter Optimization in the Classifier.

And the Pipeline to the BioGrid Interspecies Problem could be generalized into the structure represented in Figure 24.

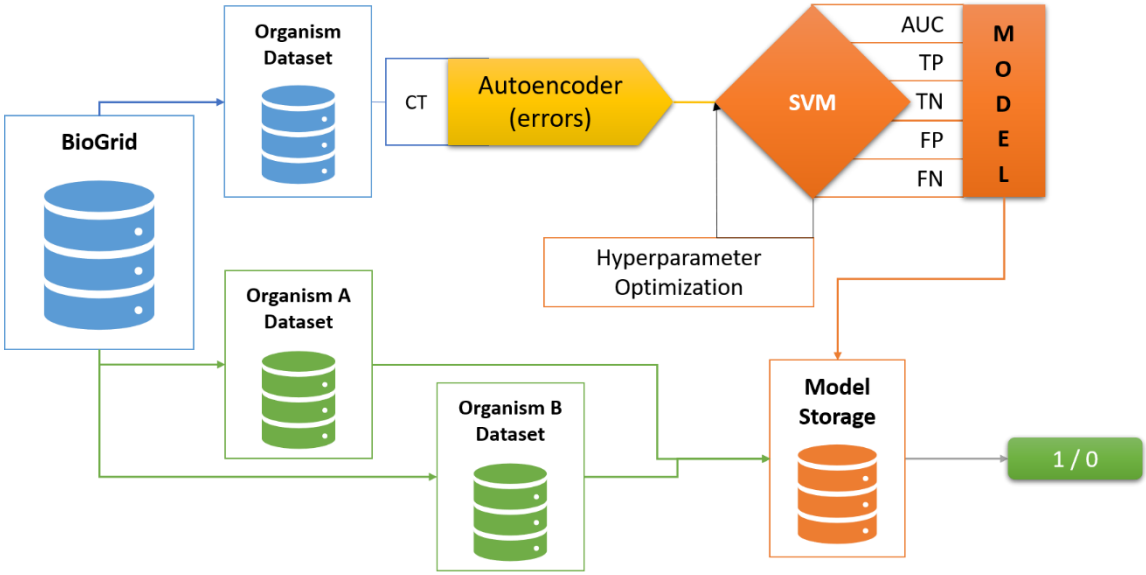

Figure 24 Final Pipeline Constructed for BioGrid Interspecies problem with the testing of different datasets to test and storage of models, applied to other organisms.

This structure allows for the classification of organisms in their relation, but also test the interactome of a new found or tested organism, in order to develop its positioning in the phylogeny tree with several organisms. In Figure 24, training organism would be Organism A and B the one tested.

However, the Pipeline allows for a single organism to be tested against all organisms, creating lists of similarities that can be used in the development of new drugs.

In a final consideration about the model Pipeline constructed, few points need to be explained through the scheme, such as:

1. Encoding Method and Area Under the Curve are immutable in the Pipeline;
2. Autoencoders and Support Vector Machines need to be optimized in the Hyperparameters search, using the specific function, and corresponding to each training organism;
3. The use of the model must be preferential in a descent level of complexity in the organisms, since low-complex organisms have more difficulties in predict higher in the hierarchy organisms.

The use of non-optimized models following these rules can lead to invalid results, creating false expectations in the laboratory experiments, and to the spending of rare resources and money.

7 Final Considerations and Future Work

In the development of this work, some interesting relationships can be found between organisms, and the research that was developed, as well as the methodologies and principles followed can be referenced in three topics.

The first, related to the objectives of this thesis, was fulfilled, as we came up with a defined Pipeline to predict protein interactions both in one organism, and interspecies, with results that equal, at least, the current literature, and that give space for the development of techniques that use the errors of Neural Networks, this case Autoencoders, as features to be introduced in a classifier. This approach is not a novelty but introduces a new look at the potential of neural networks, as clustering algorithms, with the exchange and analogies possible between them. The ensemble of classifiers can overcome problems such as dimensionality in a basic problem such as Protein Interaction prediction

The second, in the parallelism between the biochemical and bioinformatic approaches made in Chapter 3, can be a positive step forward in the openness of Bioinformatics to all the researchers in the field of biology allowing the entrance in new fields.

And for the last, the models and equations developed may help in the creation of a more complex tool, with the following characteristics / functions, as well as others thought:

1) Insertion of two protein ID's
    a) Output the existence of interactions
2) Insertion of a protein pair (or several) ID's and an Organism taxonomy ID
    a) Output the interaction prediction for each pair based on the Organism's dataset
3) Insertion of several interactions and an Organism taxonomy ID
    a) Increase the dataset of known protein interactions and retrain a model
4) Insertion of new Organism taxonomy ID and its dataset of interactions

But the development can and should be implemented in new datasets and organisms, resorting to phylogeny as an initial choice and filter factor and using high complexity level organisms as the first in line to test the similarities between organisms.

The search for interactions must be however amplified and complexified. The non-existence of a triple protein interaction database debilitates the response by bioinformatics, biochemists and pharmaceutical industries in the field of Drug Discovery [59]. These interactions could be the next strong step for computational models as they also grow in complexity.

In a final consideration, the work in the field of Protein Interactions can show where the dangers for the human body are installed, creating a map of organisms and proteins that are likely to interact with our proteins.

This information can be used in both prevention of diseases and in the Drug Development process, including new biologic macromolecules in the role of testing drugs, finding natural substitutes to medicine with high probability of secondary effects occurrence and decreasing the synthesis of man-made drugs, returning to a symbiosis between the environment and humanity, and the exploitation of some rare raw materials.

8 Bibliography

[1] N. Siddharthan, M. R. Prabu, and B. Sivasankari, "Bioinformatics in Drug Discovery a Review," no. April, pp. 10–13, 2016.

[2] S. DECKER and E. A. SAUSVILLE, "CHAPTER 28 - Drug Discovery," A. J. Atkinson, D. R. Abernethy, C. E. Daniels, R. L. Dedrick, and S. P. B. T.-P. of C. P. (Second E. Markey, Eds. Burlington: Academic Press, 2007, pp. 439–447.

[3] L. Srihari, Sriganesh; Yong, Chern Han; Wong, *Computational Prediction of protein Complexes from Protein Interaction Networks*. Morgan & Claypool Publishers, 2017.

[4] S. Dibyajyoti, E. Talha Bin, and P. Swati, "Bioinformatics:The effects on the cost of drug discovery," *Gall. Med. J.*, vol. 18, no. 1, p. 44, 2013.

[5] M. Karlgren *et al.*, "Classification of inhibitors of hepatic organic anion transporting polypeptides (OATPs): Influence of protein expression on drug-drug interactions," *J. Med. Chem.*, vol. 55, no. 10, pp. 4740–4763, 2012.

[6] U. of Coimbra, "D4 Project," 2018. .

[7] H. R. DiMasi JA, Grabowski HG, "Innovation in the pharmaceutical industry: new estimates of R&D costs," *J. Health Econ.*, 2016.

[8] E. O. Wilson, *Biodiversity*. National Academy of Sciences / Smithsonian Institution, 1988.

[9] David W. Thomas, Justin Burns, John Audette, Adam Carroll, Corey Dow-Hygelund, and Michael Hay, "Clinical Development Success Rates," *BioMedTracker*, vol. June, no. June, 2016.

[10] D. L. N. (University of W. M. M. C. (University of Wisconsin–Madison), *PRINCIPLES OF BIOCHEMISTRY*, vol. 17, no. 8. 2011.

[11] J. Zahiri, O. Yaghoubi, M. Mohammad-Noori, R. Ebrahimpour, and A. Masoudi-Nejad, "PPIevo: Protein-protein interaction prediction from PSSM based evolutionary information," *Genomics*, vol. 102, no. 4, pp. 237–242, 2013.

[12] Y. Yang and Y. Zhou, "Specific interactions for ab initio folding of protein terminal regions with secondary structures," *Proteins Struct. Funct. Genet.*, vol. 72, no. 2, pp. 793–803, 2008.

[13] Alessandro Paciaroni; E Cornicchi; A De Francesco, "Conditioning action of the environment on the protein dynamics studied through elastic neutron scattering," *Eur. Biophys. J.*, 2006.

[14] A. Bateman *et al.*, "UniProt: The universal protein knowledgebase," *Nucleic Acids Res.*, vol. 45, no. D1, pp. D158–D169, 2017.

[15] C. Vogel and E. M. Marcotte, "Insights into the regulation of protein abundance from proteomic and transcriptomic analyses," *Nat. Rev. Genet.*, vol. 13, p. 227, Mar. 2012.

[16] C. Yan, F. Wu, R. L. Jernigan, D. Dobbs, and V. Honavar, "Characterization of protein-protein interfaces," *Protein J.*, vol. 27, no. 1, pp. 59–70, 2008.

[17] L. A. Banaszynski and T. J. Wandless, "Conditional control of protein function," *Chem. Biol.*, vol. 13, no. 1, pp. 11–21, 2006.

[18] H. K. Ly, M. Sezer, N. Wisitruangsakul, J. Feng, and A. Kranich, "Surface-enhanced vibrational spectroscopy for probing transient interactions of proteins with biomimetic interfaces : electric field effects on structure , dynamics and function of cytochrome c," vol. 278, pp. 1382–1390, 2011.

[19] S. Liu, C. Liu, and L. Deng, "Machine learning approaches for protein-protein interaction hot spot prediction: Progress and comparative assessment," *Molecules*, vol. 23, no. 10, 2018.

[20] J. de Las Rivas and C. Fontanillo, "Protein-protein interactions essentials: Key concepts to building and analyzing interactome networks," *PLoS Comput. Biol.*, vol. 6, no. 6, pp. 1–8, 2010.

[21] L. T. Dang *et al.*, "Trawler Web: An online de novo motif discovery tool for next-generation sequencing datasets," *BMC Genomics*, vol. 19, no. 1, pp. 1–9, 2018.

[22] R. Kohavi and G. H. John, "Wrappers for feature subset selection," *Artif. Intell.*, vol. 97, no. 1, pp. 273–324, 1997.

[23] G. Chandrashekar and F. Sahin, "A survey on feature selection methods," *Comput. Electr. Eng.*, vol. 40, no. 1, pp. 16–28, 2014.

[24] C. D. Hu, Y. Chinenov, and T. K. Kerppola, "Visualization of interactions among bZIP and Rel family proteins in living cells using bimolecular fluorescence complementation," *Mol. Cell*, vol. 9, no. 4, pp. 789–798, 2002.
[25] O. Puig *et al.*, "The tandem affinity purification (TAP) method: A general procedure of protein complex purification," *Methods*, vol. 24, no. 3, pp. 218–229, 2001.
[26] O. Söderberg *et al.*, "Characterizing proteins and their interactions in cells and tissues using the in situ proximity ligation assay," *Methods*, vol. 45, no. 3, pp. 227–232, 2008.
[27] R. B. Free, L. A. Hazelwood, and D. R. Sibley, "Identifying novel protein-protein interactions using co-immunoprecipitation and mass spectroscopy," *Curr. Protoc. Neurosci.*, no. SUPPL. 46, pp. 1–19, 2009.
[28] Y. Li, C. Huang, L. Ding, Z. Li, Y. Pan, and X. Gao, "Deep learning in bioinformatics: Introduction, application, and perspective in the big data era," *Methods*, 2019.
[29] B. Q. Li, K. Y. Feng, L. Chen, T. Huang, and Y. D. Cai, "Prediction of Protein-Protein Interaction Sites by Random Forest Algorithm with mRMR and IFS," *PLoS One*, vol. 7, no. 8, pp. 1–10, 2012.
[30] Z. P. Liu, L. Y. Wu, Y. Wang, X. S. Zhang, and L. Chen, "Prediction of protein-RNA binding sites by a random forest method with combined features," *Bioinformatics*, vol. 26, no. 13, pp. 1616–1622, 2010.
[31] E. D. Coelho, I. N. Cruz, A. Santiago, J. L. Oliveira, A. Dourado, and J. P. Arrais, "A Sequence-Based Mesh Classifier for the Prediction of Protein-Protein Interactions," 2017.
[32] L. Lan, N. Djuric, Y. Guo, and S. Vucetic, "MS-kNN: Protein function prediction by integrating multiple data sources," *BMC Bioinformatics*, vol. 14, no. SUPPL.3, p. S8, 2013.
[33] M. R. Guarracino, I. National, and A. Nebbia, "Predicting Protein-Protein Interactions with K-Nearest Neighbors Classification Algorithm Predicting protein-protein interactions with k-Nearest Neighbors classification algorithm," no. October, 2009.
[34] J. Y. An, F. R. Meng, Z. H. You, X. Chen, G. Y. Yan, and J. P. Hu, "Improving protein–protein interactions prediction accuracy using protein evolutionary information and relevance vector machine model," *Protein Sci.*, vol. 25, pp. 1825–1833, 2016.
[35] Z. H. You, Y. K. Lei, L. Zhu, J. Xia, and B. Wang, "Prediction of protein-protein interactions from amino acid sequences with ensemble extreme learning machines and principal component analysis," *BMC Bioinformatics*, vol. 14, no. SUPPL8, pp. 1–11, 2013.
[36] Y. Zhao, J. Li, and L. Yu, "A deep learning ensemble approach for crude oil price forecasting," *Energy Econ.*, vol. 66, pp. 9–16, 2017.
[37] P. Vincent, H. Larochelle, Y. Bengio, and P. A. Manzagol, "Extracting and composing robust features with denoising autoencoders," *Proc. 25th Int. Conf. Mach. Learn.*, no. July, pp. 1096–1103, 2008.
[38] X. Qiu, L. Zhang, Y. Ren, P. Suganthan, and G. Amaratunga, "Ensemble deep learning for regression and time series forecasting," *IEEE SSCI 2014 - 2014 IEEE Symp. Ser. Comput. Intell. - CIEL 2014 2014 IEEE Symp. Comput. Intell. Ensemble Learn. Proc.*, pp. 1–6, 2014.
[39] G. P. Way and C. S. Greene, "Extracting a biologically relevant latent space from cancer transcriptomes with variational autoencoders," *Pacific Symp. Biocomput.*, vol. 0, no. 212669, pp. 80–95, 2018.
[40] J. W. Liu, G. H. Chi, Z. Y. Liu, L. Yuan, H. E. Li, and X. L. Luo, "Predicting protein structural classes with autoencoder neural networks," *2013 25th Chinese Control Decis. Conf. CCDC 2013*, pp. 1894–1899, 2013.
[41] J. Albuquerque and C. Pereira, "Analysis of Autoencoders for Feature Representation of Protein Sequences 3 Autoencoders 1 Introduction 4 Experiments and Dataset 2 Protein Sequences Feature Extraction 5 Results and Discussion 6 Conclusion Acknowledgments," pp. 4–5.
[42] M. H. Beale, M. T. Hagan, and H. B. Demuth, "Deep Learning Toolbox ™ Reference How to Contact MathWorks," 2019.
[43] K. Nakai and A. Patil, "Database update HitPredict version 4 : comprehensive reliability scoring of physical protein – protein interactions from more than 100 species," pp. 1–10, 2015.
[44] C. Stark, "BioGRID: a general repository for interaction datasets," *Nucleic Acids Res.*, vol.

34, no. 90001, pp. D535–D539, 2006.
[45] U. S. N. L. of M. National Center for Biotechnology Information, “NCBI - BLAST.” [Online]. Available: https://blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Web&PAGE_TYPE=BlastDocs&DOC_TYPE=BlastHelp.
[46] P. Du and Y. Li, “Prediction of protein submitochondria locations by hybridizing pseudo-amino acid composition with various physicochemical features of segmented sequence,” *BMC Bioinformatics*, vol. 7, pp. 1–8, 2006.
[47] Y. Xu, X. Wen, X. Shao, N. Deng, and K. Chou, “iHyd-PseAAC : Predicting Hydroxyproline and Hydroxylysine in Proteins by Incorporating Dipeptide Position-Specific Propensity into Pseudo Amino Acid Composition,” pp. 7594–7610, 2014.
[48] J. Shen *et al.*, “Predicting protein-protein interactions based only on sequences information,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 104, no. 11, pp. 4337–4341, 2007.
[49] Q. SONG, “A NOTE ON FUZZY TIME SERIES MODEL SELECTION WITH SAMPLE AUTOCORRELATION FUNCTIONS,” *Cybern. Syst.*, vol. 34, no. 2, pp. 93–107, Mar. 2003.
[50] M. Ganapathiraju, D. Weisser, R. Rosenfeld, J. Carbonell, R. Reddy, and S. Francisco, “Comparative n-gram analysis of whole-genome protein sequences Conference on Human Language Technology,” pp. 76–81, 2002.
[51] F. A. Gers and F. Cummins, “Learning to Forget: Continual prediction with LSTM,” pp. 1–19, 1999.
[52] T. Sun, B. Zhou, L. Lai, and J. Pei, “Sequence-based prediction of protein protein interaction using a deep-learning algorithm,” *BMC Bioinformatics*, vol. 18, no. 1, pp. 1–8, 2017.
[53] Y. Guo *et al.*, “PRED _ PPI : a server for predicting protein-protein interactions based on sequence data with probability assignment,” pp. 1–7, 2010.
[54] X.-Y. Pan, Y.-N. Zhang, and H.-B. Shen, “Large-Scale Prediction of Human Protein−Protein Interactions from Amino Acid Sequence Based on Latent Topic Features,” *J. Proteome Res.*, vol. 9, no. 10, pp. 4992–5001, Oct. 2010.
[55] Y.-N. Zhang, X.-Y. Pan, Y. Huang, and H.-B. Shen, “Adaptive compressive learning for prediction of protein–protein interactions from primary sequence,” *J. Theor. Biol.*, vol. 283, no. 1, pp. 44–52, 2011.
[56] Z. You, S. Li, X. Gao, X. Luo, and Z. Ji, “Large-Scale Protein-Protein Interactions Detection by Integrating Big Biosensing Data with Computational Model,” vol. 2014, pp. 28–30, 2014.
[57] Z. You, J. Yu, L. Zhu, S. Li, and Z. Wen, “A MapReduce based parallel SVM for large-scale predicting protein-protein interactions,” *Neurocomputing*, 2014.
[58] S. Hedges, S. B., Marin, J., Suleski, M., Paymer M., Kumar, “Timetree of Life (2015) Download (>50,000 species).” 2015.
[59] Ø. Halskau *et al.*, “Three-way Interaction between 14-3-3 Proteins , the N-terminal Region of Tyrosine Hydroxylase , and Negatively Charged Membranes * □,” vol. 284, no. 47, pp. 32758–32769, 2009.

Annexes

Annex 1 - Table of Physic-chemical properties of amino acids

aa.	H_1	H_2	NCI	P_1	P_2	SASA	V
A	0.620	-0.500	0.007	8.100	0.046	1.181	27.500
C	0.290	-1	-0.037	5.500	0.128	1.461	44.600
D	-0.900	3	-0.024	13.000	0.105	1.587	40.000
E	-0.740	3	0.007	12.300	0.151	1.862	62.000
F	1.190	-2.500	0.038	5.200	0.290	2.228	115.500
G	0.480	0	0.179	9.000	0.000	0.881	0.000
H	-0.400	-0.500	-0.011	10.400	0.230	2.025	79.000
I	1.380	-1.800	0.022	5.200	0.186	1.810	93.500
K	-1.500	3	0.018	11.300	0.219	2.258	100.000
L	1.060	-1.800	0.052	4.900	0.186	1.931	93.500
M	0.640	-1.300	0.003	5.700	0.221	2.034	94.100
N	-0.780	2	0.005	11.600	0.134	1.655	58.700
P	0.120	0	0.240	8.000	0.131	1.468	41.900
Q	-0.850	0.200	0.049	10.500	0.180	1.932	80.700
R	-2.530	3	0.044	10.500	0.291	2.56	105.000
S	-0.180	0.300	0.005	9.200	0.062	1.298	29.300
T	-0.050	-0.400	0.003	8.600	0.108	1.525	51.300
V	1.080	-1.500	0.057	5.900	0.140	1.645	71.500
W	0.810	-3.400	0.038	5.400	0.409	2.663	145.500
Y	0.260	-2.300	117.300	6.200	0.298	2.368	0.024

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;

Annex 2 - Table of Amino acid classification for CT encoding

Number	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

Annex 3 - Table of 26 Organisms Processed

UniProt ID	*Scientific Name*	*Common Name*	*Nr of Positive Interactions*
10090	Mus Musculus	Mouse	30 000
10116	Rattus Norvegicus	Rat	6 637
10298	Human herpesvirus 1	Human Herpes Virus	330
10359	Human cytomegalovirus	Human Herpes Virus 5	285
10376	Epstein-Barr virus	HHV 4	1 074
11676	HIV		3 471
227321	Emericella nidulans	Aspergillus nidulans	65
237561	Candida albicans	Yeast	392
284812	Schizosaccharomyces pombe	Fission yeast	30 000
333760	Human papillomavirus		145
3702	Arabidopsis thaliana	Arabidopsia	30 000
37296	Human herpesvirus 8		319
3847	Glycine max	Soybean; Glycine hispida	36
39947	Oryza sativa subsp. japonica	Rice	49
511145	Escherichia coli str. K-12 substr. MG1655		96
559292	Saccharomyces cerevisiae	Baker's yeast	30 000
6239	Caenorhabditis elegans		26 483
7227	Drosophila melanogaster	Fruit fly	30 000
7955	Danio rerio		423
83332	Mycobacterium tuberculosis		41
8355	Xenopus laevis		1 212
9031	Gallus gallus		488
9606	Homo Sapiens		41 273
9823	Sus scrofa		46
9913	Bos taurus		393
9986	Oryctolagus cuniculus	Rabbit	203

Annex 4 – Amino Acid Code

Amino acid	***Three letter code***	***One letter code***
Alanine	ala	A
Arginine	arg	R
Asparagine	asn	N
Aspartic acid	asp	D
Asparagine or aspartic acid	asx	B
Cysteine	cys	C
Glutamic acid	glu	E
Glutamine	gln	Q
Glutamine or glutamic acid	glx	Z
Glycine	gly	G
Histidine	his	H
Isoleucine	ile	I
Leucine	leu	L
Lysine	lys	K
Methionine	met	M
Phenylalanine	phe	F
Proline	pro	P
Serine	ser	S
Threonine	thr	T
Tryptophan	trp	W
Tyrosine	tyr	Y
Valine	val	V

Annex 5 - Overlapping matrix between organisms' datasets

	10090	10116	10298	10359	10376	11676	227321	237561	284812	333760	3702	37296	3847	39947	511145	559292	6239	7227	7955	83332	8355	9031	9606	9823	9913	9986
10090	100,00%	0,02%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,07%	0,00%	0,01%	0,00%
10116	0,12%	100,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,03%	0,00%	0,02%	0,00%	0,01%	0,00%
10298	0,00%	0,00%	100,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
10359	0,00%	0,00%	0,00%	100,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
10376	0,00%	0,00%	0,00%	0,00%	100,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
11676	0,00%	0,00%	0,00%	0,00%	0,00%	100,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
227321	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	100,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
237561	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	100,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
284812	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	100,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
333760	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	100,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
3702	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	100,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
37296	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	100,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
3847	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	100,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
39947	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	100,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
511145	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	100,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
559292	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	100,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
6239	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	100,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
7227	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	100,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
7955	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	100,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
83332	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	100,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
8355	0,00%	0,17%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	100,00%	0,00%	0,00%	0,00%	0,00%	0,00%
9031	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	100,00%	0,00%	0,00%	0,00%	0,00%
9606	0,07%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	100,00%	0,00%	0,00%	0,00%
9823	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	100,00%	0,00%	0,00%
9913	0,77%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,51%	0,00%	100,00%	0,00%
9986	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	100,00%

Annex 6 - AUC of Interspecies testing with K-NN

	10090	10116	10238	10359	10376	11676	227321	237561	284812	333760	3702	37296	3847	39947	511145	559292	6239	7227	7955	83332	8355	9031	9606	9823	9913	9986
10090	0.849	0.529	0.693	0.807	0.897	0.660	0.609	0.901	0.678	0.907	0.568	0.743	0.757	0.780	0.893	0.838	0.808	0.558	0.796	0.516	0.702	0.879	0.501	0.728	0.635	0.525
10116	0.677	0.523	0.841	0.863	0.655	0.683	0.578	0.510	0.904	0.507	0.744	0.914	0.913	0.746	0.639	0.621	0.553	0.706	0.649	0.810	0.502	0.614	0.761	0.503	0.675	
10238	0.828	0.864	0.904	0.774	0.500	0.506	0.641	0.811	0.593	0.972	0.607	0.788	0.716	0.643	0.530	0.658	0.867	0.564	0.848	0.752	0.504	0.785	0.581	0.583		
10359	0.837	0.953	0.982	0.790	0.501	0.612	0.501	0.818	0.570	0.978	0.753	0.792	0.503	0.639	0.592	0.509	0.879	0.523	0.780	0.562	0.662	0.502	0.577	0.797		
10376	0.960	0.800	0.985	0.664	0.621	0.587	0.535	0.928	0.592	0.929	0.696	0.814	0.746	0.546	0.524	0.573	0.851	0.542	0.680	0.785	0.510	0.708	0.605	0.792		
11676	0.785	0.849	0.911	0.888	0.783	0.796	0.552	0.891	0.609	0.626	0.912	0.641	0.765	0.594	0.528	0.650	0.576	0.510	0.512	0.811	0.570	0.726	0.538	0.812		
227321	0.551	0.523	0.691	0.510	0.608	0.695	0.557	0.629	0.596	0.511	0.529	0.740	0.766	0.583	0.547	0.539	0.589	0.571	0.609	0.640	0.527	0.687	0.569	0.715		
237561	0.584	0.713	0.612	0.660	0.622	0.825	0.628	0.699	0.505	0.818	0.714	0.739	0.739	0.580	0.618	0.681	0.559	0.597	0.580	0.827	0.597	0.881	0.503	0.732		
284812	0.564	0.778	0.541	0.612	0.694	0.852	0.753	0.613	0.526	0.616	0.570	0.553	0.661	0.525	0.622	0.544	0.539	0.523	0.592	0.525	0.668	0.616	0.622			
333760	0.874	0.914	0.528	0.576	0.800	0.521	0.922	0.562	0.798	0.677	0.891	0.561	0.585	0.545	0.539	0.694	0.531	0.540	0.653	0.532	0.560	0.551	0.705			
3702	0.633	0.609	0.526	0.569	0.634	0.608	0.899	0.934	0.810	0.689	0.875	0.715	0.595	0.687	0.751	0.585	0.586	0.553	0.739	0.504	0.885	0.579	0.682			
37296	0.898	0.808	0.709	0.512	0.694	0.543	0.838	0.577	0.789	0.779	0.865	0.687	0.589	0.774	0.654	0.748	0.526	0.695	0.596	0.510	0.503	0.522	0.694			
3847	0.893	0.956	0.743	0.671	0.838	0.613	0.985	0.546	0.945	0.684	0.934	0.779	0.650	0.582	0.672	0.876	0.679	0.812	0.879	0.571	0.779	0.734	0.574			
39947	0.886	0.925	0.544	0.619	0.644	0.553	0.908	0.667	0.784	0.633	0.969	0.581	0.600	0.578	0.506	0.801	0.618	0.676	0.682	0.642	0.536	0.588	0.789			
511145	0.949	0.808	0.775	0.669	0.755	0.724	0.836	0.605	0.990	0.680	0.705	0.638	0.626	0.651	0.656	0.846	0.617	0.746	0.807	0.630	0.797	0.735	0.598			
559292	0.691	0.512	0.526	0.544	0.538	0.641	0.669	0.875	0.745	0.517	0.502	0.607	0.851	0.547	0.597	0.658	0.566	0.648	0.615	0.543	0.525	0.508	0.652			
6239	0.861	0.977	0.591	0.588	0.563	0.609	0.814	0.511	0.711	0.794	0.830	0.538	0.588	0.576	0.501	0.641	0.576	0.771	0.626	0.503	0.561	0.561	0.528			
7227	0.832	0.706	0.594	0.555	0.678	0.697	0.939	0.543	0.589	0.636	0.662	0.528	0.981	0.584	0.949	0.665	0.596	0.560	0.505	0.525	0.591	0.626	0.819			
7955	0.850	0.794	0.616	0.575	0.543	0.527	0.814	0.598	0.671	0.733	0.924	0.660	0.525	0.552	0.554	0.676	0.596	0.569	0.626	0.648	0.517	0.508	0.604			
83332	0.762	0.925	0.503	0.581	0.697	0.693	0.882	0.607	0.779	0.642	0.894	0.555	0.571	0.563	0.508	0.731	0.642	0.616	0.604	0.586	0.562	0.519	0.796			
8355	0.830	0.762	0.536	0.639	0.650	0.881	0.601	0.966	0.943	0.818	0.842	0.773	0.609	0.638	0.616	0.866	0.588	0.773	0.748	0.589	0.775	0.729	0.745			
9031	0.814	0.617	0.574	0.893	0.778	0.816	0.634	0.943	0.589	0.763	0.748	0.865	0.500	0.662	0.523	0.814	0.546	0.722	0.715	0.635	0.751	0.716	0.573			
9606	0.706	0.555	0.588	0.543	0.527	0.693	0.882	0.607	0.779	0.642	0.894	0.555	0.571	0.563	0.508	0.731	0.642	0.616	0.604	0.586	0.562	0.519	0.796			
9823	0.879	0.555	0.588	0.543	0.527	0.693	0.882	0.607	0.779	0.642	0.894	0.555	0.571	0.563	0.508	0.731	0.642	0.616	0.604	0.586	0.562	0.519	0.796			
9913	0.826	0.958	0.783	0.503	0.787	0.505	0.849	0.516	0.830	0.646	0.751	0.805	0.532	0.608	0.517	0.638	0.590	0.692	0.705	0.663	0.755	0.558	0.556			
9986	0.776	0.713	0.807	0.660	0.586	0.520	0.980	0.654	0.695	0.837	0.606	0.602	0.565	0.628	0.500	0.631	0.635	0.528	0.555	0.505	0.598	0.618	0.777			

Annex 7 - AUC of Interspecies testing with SVM

	10090	10116	10238	10359	10376	11676	227321	237561	284812	333760	3702	37296	3847	39947	511145	559292	6239	7227	7955	83332	8355	9031	9606	9823	9913	9986
10090	0,8629	0,598944	0,7278	0,992601	0,934074	0,855852	0,773243	0,902497	0,861004	1	0,7564	0,884153	0,791667	0,851563	0,947778	0,8926	0,820108	0,612	0,859326	0,5625	0,819966	0,947152	0,5135	0,902222	0,77092	0,711238
10116	0,633347	0,8024	0,6521	0,849254	0,921546	0,840854	0,716553	0,532699	0,524477	0,989136	0,554284	0,837177	0,916667	0,914063	0,812222	0,511575	0,796057	0,534841	0,777836	0,5	0,824898	0,633755	0,505104	0,884444	0,630676	0,689126
10298	0,577804	0,587936	0,9839	1	0,997778	0,894963	0,693878	0,692628	0,767171	1	0,546826	0,997662	0,75	0,957031	0,803333	0,509951	0,707983	0,517602	0,943548	0,625	0,864868	0,903883	0,513955	0,897778	0,701375	0,738554
10359	0,591706	0,592238	0,9902	0,9815	0,985354	0,853413	0,648526	0,638823	0,540957	1	0,599685	0,998618	0,875	0,964844	0,621111	0,552903	0,764473	0,509049	0,900234	0,625	0,809503	0,724652	0,508382	0,577778	0,641692	0,869667
10376	0,571975	0,59894	0,9806	0,998679	0,9949	0,840629	0,680272	0,736326	0,587811	1	0,540394	0,999787	0,854167	0,960938	0,848889	0,506943	0,723262	0,514259	0,894186	0,5	0,859899	0,864964	0,500526	0,835556	0,747122	0,797086
11676	0,600949	0,525622	0,9788	0,984542	0,969673	0,9187	0,902494	0,805291	0,50604	1	0,518851	0,664895	0,979167	0,542969	0,795556	0,552351	0,653236	0,511071	0,587864	0,5	0,512914	0,932852	0,514592	0,911111	0,545724	0,895552
227321	0,500528	0,533478	0,6047	0,602061	0,805232	0,662316	0,6429	0,887337	0,721214	0,511111	0,542962	0,677968	0,583333	0,769531	0,794444	0,527794	0,562671	0,500567	0,519251	0,5625	0,564293	0,756259	0,501204	0,853333	0,572841	0,791363
237561	0,609342	0,619369	0,7172	0,745541	0,727669	0,530448	0,528345	0,9444	0,820991	0,888889	0,53745	0,936019	0,743056	0,757813	0,775556	0,515299	0,690155	0,509248	0,617391	0,5625	0,705348	0,999285	0,523317	0,937778	0,638867	0,8564
284812	0,603262	0,575344	0,626	0,581054	0,830268	0,566926	0,539683	0,752973	0,8761	0,885432	0,528489	0,661388	0,770833	0,53125	0,591111	0,512071	0,564281	0,50946	0,725676	0,5625	0,695988	0,762599	0,527039	0,848889	0,522986	0,669355
333760	0,531701	0,569614	0,9281	0,975294	0,966654	0,652423	0,569161	0,811534	0,646886	1,0000	0,503617	0,864598	0,715278	0,933594	0,58	0,589384	0,516952	0,53248	0,859912	0,5625	0,538773	0,660373	0,51622	0,76	0,512746	0,8826674
3702	0,8254	0,551226	0,5035	0,731537	0,515025	0,657093	0,575964	0,670927	0,550496	0,910123	0,9955	0,863429	0,805556	0,917969	0,865556	0,6581	0,6885	0,8109	0,608741	0,5625	0,558686	0,8614	0,5992	0,92	0,556034	0,726327
37296	0,616495	0,613932	0,9496	0,991941	0,982748	0,796009	0,680272	0,502081	0,570187	1	0,569292	0,9833	0,847222	0,871094	0,705556	0,540587	0,779321	0,519617	0,862838	0,5625	0,78471	0,78148	0,520386	0,6	0,671139	0,843132
3847	0,571531	0,671773	0,9975	1	0,996751	0,823351	0,503401	0,865636	0,531696	1	0,583809	0,97157	0,8000	0,976563	0,813333	0,507213	0,707595	0,517449	0,886837	0,71875	0,830407	0,881901	0,507657	0,968889	0,778123	0,656738
39947	0,642257	0,609341	0,9859	0,998679	0,97805	0,905557	0,739229	0,931629	0,81135	1	0,583258	0,99543	0,840278	0,894531	0,7333	0,552765	0,779291	0,506676	0,869472	0,5625	0,830858	0,95997	0,520852	0,96	0,816891	0,598855
511145	0,97	0,514642	0,7712	0,738142	0,649656	0,645943	0,553288	0,648633	0,52868	0,857778	0,9324	0,832395	0,5	0,570313	0,53	0,9637	0,642854	0,7815	0,806126	0,5	0,546617	0,551946	0,6389	0,555556	0,581244	0,685484
559292	0,646382	0,550491	0,98025	0,954684	0,993741	0,752264	0,698413	0,640309	0,562761	0,99358	0,588958	0,892231	0,902778	0,832031	0,715556	0,550099	0,7055	0,529664	0,833572	0,625	0,840585	0,719145	0,506269	0,604444	0,634348	0,694199
6239	0,5439	0,516159	0,5819	0,947417	0,865502	0,662581	0,582766	0,689061	0,768181	1	0,6428	0,589223	0,701389	0,828125	0,687778	0,9886	0,530986	0,9725	0,727693	0,5	0,56945	0,641168	0,6723	0,68	0,523409	0,837669
7227	0,534942	0,514446	0,7796	0,937772	0,955851	0,557921	0,548753	0,530618	0,618164	0,955556	0,538726	0,843979	0,798611	0,972656	0,666667	0,644583	0,532716	0,536464	0,8553	0,625	0,679737	0,666204	0,52114	0,577778	0,569239	0,593913
7955	0,544668	0,554346	0,8851	0,957062	0,963807	0,612547	0,60771	0,726813	0,743998	1	0,527966	0,798703	0,708333	0,917969	0,62	0,616836	0,533216	0,542702	0,756243	0,5000	0,506612	0,587533	0,503364	0,751111	0,503496	0,871488
83332	0,56379	0,639389	0,9891	1	0,997479	0,829433	0,630385	0,739075	0,54811	1	0,572922	0,998725	0,840278	1	0,856667	0,500098	0,71609	0,520902	0,918054	0,5625	0,8000	0,907724	0,511701	0,888889	0,77205	0,758845
8355	0,63658	0,600272	0,9888	0,977408	0,997494	0,801045	0,662132	0,932818	0,865785	0,918519	0,559845	0,994792	0,791667	0,882813	0,957778	0,520672	0,761592	0,528016	0,866285	0,5625	0,817234	0,9667	0,519374	0,937778	0,788645	0,55307
9031	0,8703	0,527203	0,767	0,924032	0,911779	0,512074	0,521542	0,661712	0,846884	0,800988	0,9826	0,595813	0,777778	0,851563	0,722222	0,9568	0,537596	0,5161	0,812955	0,8125	0,730291	0,714887	0,9704	0,884889	0,604053	0,652445
9606	0,632015	0,612986	0,9847	0,99379	0,99978	0,916276	0,662132	0,821046	0,618744	1	0,578954	1	0,875	0,875	0,838889	0,549117	0,80913	0,500227	0,852367	0,5	0,817347	0,859235	0,507681	0,8000	0,782713	0,699011
9823	0,568571	0,599724	0,9846	1	0,97898	0,833358	0,675737	0,873365	0,516635	1	0,587098	0,988734	0,75	0,898438	0,961111	0,59696	0,699003	0,512689	0,762812	0,5	0,802454	0,89236	0,503555	0,942222	0,7530	0,633455
9986	0,507734	0,568913	0,5215	0,960497	0,88958	0,814711	0,718821	0,707491	0,622437	1	0,528849	0,877458	0,909722	0,714844	0,704444	0,535193	0,674825	0,533974	0,718392	0,5625	0,672794	0,746726	0,501781	0,515556	0,641974	0,8947