# Accelerated Transfer Learning for Protein-Protein Interaction Prediction

By

Bradley Dennis Barnes

A thesis submitted to the Faculty of Graduate Studies and Research in partial fulfillment of the requirements for the degree of

**Masters of Applied Science**

**in Biomedical Engineering with Specialization in Data Science**

Ottawa-Carleton Institute for Biomedical Engineering

Department of Systems and Computer Engineering
Carleton University
Ottawa, Ontario, Canada
January 2018

# Abstract

This thesis explores issues arising when one attempts to predict protein-protein interactions (PPI) involving multiple species using the Protein-protein Interaction Prediction Engine (PIPE) method. In cross-species predictions, where one predicts PPI in a target species given known PPI in a different training species, we showed that prediction performance is inversely correlated to the evolutionary distance between training and target species. With a change in the score calculation, we improved the area under the precision-recall curve by 45% when using seven well-studied species to predict an eighth.

In inter-species predictions, one attempts to predict interactions between proteins arising from two different species, such as a host and a pathogen. For the first time, we have shown that PIPE is able to predict such inter-species PPI by predicting 229 novel PPI between HIV and human at an estimated precision of 82% (100:1 class imbalance).

Lastly, by modifying a main data structure of PIPE, we also improved the speed of the PIPE algorithm by a factor of 53x when predicting *H. sapiens* PPI. Using the methods developed in this thesis, we have predicted all possible PPI between soybean and the Soybean Cyst Nematode pathogen. Collaborators at Agriculture and Agri-Food Canada will be pursuing and validating these predictions as they seek to combat this costly pest.

# Acknowledgements

I would like to thank my supervisor, James Green for his support, patience and guidance throughout this experience. I am grateful for the opportunity he provided me with to pursue this research and for all the knowledge he shared with me throughout this time.

I would like to thank all the members of the Carleton University Bioinformatics Group for their collaboration and advice throughout my work. Specifically, I would like to thank Andrew Schoenrock for helping to get me started with PIPE.

I would like to thank Bahram Samanfar and Elroy Cober from Agriculture Canada for providing information and guidance on the SCN problem that motivated much of this work. I would also like to thank Agriculture Canada for the financial support they provided to support this work.

I am grateful to the lab community for the invaluable friendships and support system I formed through this environment. I would also like to thank my friends, my parents, and my partner, Melissa Matis, for their continual love and support throughout this process.

# Statement of Originality

This thesis presents the work of the author, under the supervision of Dr. James R. Green. This work was completed at Carleton University for the degree Master of Applied Science in Biomedical Engineering with specialization in Data Science. Some of these results have been presented in a conference publication:

B. Barnes, M. Karimloo, A. Schoenrock, D. Burnside, E. Cassol, A. Wong, F. Dehne, A. Golshani, J.R. Green, "Predicting Novel Protein-Protein Interactions Between the HIV-1 Virus and Homo Sapiens," in Student Conference (ISC), 2016 IEEE EMBS International, 2016, pp. 1–4.

This conference paper describes the prediction of HIV-Human interactions. The results of which form a portion of Chapter 5 and Sections 2.3.2 and 6.3.2. This paper was presented by the author at the 2016 IEEE EMBS International Student Conference (IEEE ISC 2016) for which the author received 3rd prize in the student paper competition. This author did the entirety of the technical work and wrote the full conference paper. The other authors provided edits and recommendations for analysis, in addition to making significant contributions to PIPE that enabled this study.

The author also contributed towards a second publication on the prediction of human-Zika protein-protein interactions. Although this paper was informed by the research findings of the present thesis, no text or results from this paper are included in the thesis document.

T. Kazmirchuk, K. Dick, D.J. Burnside, B. Barne, H. Motesharei, M. Hajikarimlou, K. Omidi, D. Ahmed, A. Low, C. Lettl, M. Hooshyar, A. Schoenrock, S. Pitre, M. Babu, E. Cassol, B. Samanfar, A. Wong, F. Dehne, J.R. Green, A. Golshani, 2017, "Designing Anti-Zika Virus Peptides Derived from Predicted Human-Zika Virus Protein-Protein Interactions," Computational Biology and Chemistry, 71:180-197 (doi:10.1016/j.compbiolchem.2017.10.011).

# Table of Contents


**ABSTRACT** ..... **II**

**ACKNOWLEDGEMENTS** ..... **III**

**STATEMENT OF ORIGINALITY** ..... **IV**

**TABLE OF CONTENTS** ..... **V**

**LIST OF TABLES** ..... **VII**

**LIST OF FIGURES** ..... **VIII**

**LIST OF ABBREVIATIONS** ..... **X**

**1 INTRODUCTION** ..... **1**

1.1 INTRODUCTION ..... 1
1.2 MOTIVATION ..... 3
1.3 PROBLEM STATEMENT ..... 5
1.4 ORGANIZATION OF THESIS ..... 6

**2 BACKGROUND AND LITERATURE REVIEW** ..... **7**

2.1 BACKGROUND BIOLOGY ..... 7
2.2 PATTERN CLASSIFICATION OVERVIEW ..... 8
2.3 PROTEIN-PROTEIN INTERACTION PREDICTION METHODS ..... 13
*2.3.1 Protein-Protein Interaction Prediction from Sequence* ..... *15*
*2.3.2 Intra-, Inter-, and Cross-Species PPI Prediction* ..... *19*
*2.3.3 The Protein-Protein Interaction Prediction Engine (PIPE)* ..... *21*
*2.3.4 Summary of the State of the Art* ..... *31*

**3 COMPUTATIONAL IMPROVEMENT OF PIPE** ..... **33**

3.1 INTRODUCTION ..... 33
3.2 PIPE ALGORITHM ..... 34
*3.2.1 Preprocessing and Input Data Representation* ..... *34*
*3.2.2 PIPE Score Generation Algorithm* ..... *36*
3.3 MODIFIED PIPE ALGORITHM ..... 39
*3.3.1 Modified Input Data Representation* ..... *41*
*3.3.2 Modified PIPE Score Generation Algorithm* ..... *44*
3.4 BENCHMARKING METHODOLOGY AND DATA ..... 46
3.5 RESULTS AND DISCUSSION ..... 50
3.6 CONCLUSIONS ..... 53

**4 CROSS-SPECIES PPI PREDICTION** ..... **55**

4.1 INTRODUCTION ..... 55
4.2 METHODS ..... 56
*4.2.1 Training Data* ..... *56*
*4.2.2 Normalization Factor Change* ..... *57*
*4.2.3 Evolutionary Distance Relation to Classification Performance* ..... *61*
4.3 RESULTS ..... 66
*4.3.1 Normalization Factor Accuracy Difference* ..... *66*
*4.3.2 Evolutionary Distance Relation to Accuracy* ..... *72*

4.4 DISCUSSION ..... 76
4.5 CONCLUSION ..... 78

**5 INTER-SPECIES PPI PREDICTION ..... 80**

5.1 INTRODUCTION ..... 80
5.2 METHODS ..... 81
*5.2.1 HIV-1 – Human Prediction ..... 81*
*5.2.2 SCN – Soybean PPI Prediction ..... 83*
*5.2.3 Human – Soybean PPI Prediction ..... 84*
5.3 RESULTS ..... 85
*5.3.1 HIV-1 – Human PPI Prediction ..... 85*
*5.3.2 SCN – Soybean and Human – Soybean PPI Prediction ..... 89*
5.4 DISCUSSION ..... 89
*5.4.1 HIV-1 – Human ..... 89*
*5.4.2 Under-studied Viruses ..... 91*
*5.4.3 SCN – Soybean ..... 91*
5.5 CONCLUSION ..... 92

**6 THESIS SUMMARY AND FUTURE RECOMMENDATIONS ..... 94**

6.1 CONCLUSIONS ..... 94
6.2 SUMMARY OF CONTRIBUTIONS ..... 95
6.3 RECOMMENDATIONS FOR FUTURE WORK ..... 98
*6.3.1 Computational Improvement ..... 98*
*6.3.2 Cross- and Inter- Species PPI Prediction ..... 99*
*6.3.3 General PIPE Score Improvement ..... 102*

**REFERENCES ..... 104**

**APPENDIX A: CROSS-SPECIES EXPERIMENTS ..... 113**

# List of Tables

TABLE 1: INTRA-SPECIES PPI PREDICTION TEST DATA FOR COMPARING NEW AND OLD PIPE IMPLEMENTATION ALGORITHMS. ..... 47
TABLE 2: INTER-SPECIES PREDICTIONS COMPLETED WITH PIPE AT THE BEHEST OF AGRICULTURE CANADA. ..... 48
TABLE 3: INTRA-SPECIES TIME COMPARISON TESTS FOR NEW AND OLD PIPE ALGORITHM IMPLEMENTATION. ..... 51
TABLE 4: INTER-SPECIES TIME COMPARISON TESTS FOR NEW AND OLD PIPE ALGORITHM IMPLEMENTATION. ..... 52
TABLE 5: NUMBER OF INTRA-SPECIES PPI FOR THE DIFFERENT SPECIES USED ..... 57
TABLE 6: HYPOTHETICAL PERFORMANCE RANK RESULTING FROM USING 8 DIFFERENT TRAINING SPECIES TO PREDICT *H. SAPIENS* PPI. ..... 64
TABLE 7: PAIRED T-TEST RESULTS COMPARING TWO METHODS OF NORMALIZATION FOR EACH PERFORMANCE METRIC FOR THE SINGLE TRAINING SPECIES CROSS-SPECIES PREDICTION. ..... 68
TABLE 8: PERFORMANCE METRICS FOR CROSS-SPECIES PPI PREDICTION IN DIFFERENT TEST SPECIES. ..... 68
TABLE 9: CROSS-SPECIES PERFORMANCE METRICS WHEN USING MULTIPLE TRAINING SPECIES. ..... 71
TABLE 10: PAIRED T-TEST RESULTS COMPARING TWO METHODS OF NORMALIZATION FOR EACH PERFORMANCE METRIC FOR THE MULTIPLE TRAINING SPECIES CROSS-SPECIES PREDICTION. ..... 71
TABLE 11: RANKS AND VALUES OF AU-PRC FOR PREDICTING *H. SAPIENS* FROM OTHER SPECIES. ..... 74
TABLE 12: RANKS AND VALUES OF PRECISION AT 25% TPR FOR PREDICTING *H. SAPIENS* FROM OTHER SPECIES. HERE, RANK_EVO REFLECTS RELATIVE DISTANCE (BY NUMBER OF COMMON ANCESTORS), BETWEEN THE TRAINING SPECIES AND *H. SAPIENS*. ..... 74
TABLE 13: STATISTICAL TESTS FOR RANK CORRELATION BETWEEN THE EXPECTED RANK (BASED ON EVOLUTIONARY DISTANCE) AND THE ACTUAL RANK OF THE AU-PRC AT 10:1 CI FOR EACH TEST-SPECIES. ..... 75
TABLE 14: STATISTICAL TESTS FOR RANK CORRELATION BETWEEN THE EXPECTED RANK (BASED ON EVOLUTIONARY DISTANCE) AND THE ACTUAL RANK OF THE PRECISION AT 25% TPR AT 10:1 CI (CI DOES NOT AFFECT ORDER FOR THIS METRIC). ..... 75
TABLE 15: MODIFIED TAU-B TEST TO COMBINE ALL DATA SOURCES INTO ONE EXPERIMENT. ..... 76
TABLE 16: DATA SUMMARY OF THE HIV-1 INTERACTIONS DATABASE AND SWISS-PROT HUMAN. ..... 82
TABLE 17: PRECISION ESTIMATE FOR THE 229 NOVEL PPI PREDICTIONS AT DIFFERENT CLASS IMBALANCES AT THE FINAL DECISION THRESHOLD. ..... 89
TABLE 18: SINGLE-SPECIES CROSS-SPECIES PREDICTION RESULTS FOR TWO PERFORMANCE METRICS (EVALUATED AT 10:1 CI) FOR EVERY PAIR OF TRAINING AND TESTING SPECIES. ..... 129
TABLE 19: FULL RANKINGS FOR EACH PERFORMANCE METRIC OF CROSS-SPECIES EXPERIMENTS (CONTROLLING FOR TRAINING SET SIZE). ..... 139

# List of Figures


FIGURE 1: PREVALENCE OF SCN ACROSS COUNTIES IN CANADA AND THE US (REPRODUCED FROM [8]). ..... 2
FIGURE 2: HIGH-LEVEL OVERVIEW OF CROSS-SPECIES AND INTER-SPECIES PPI PREDICTION ..... 5
FIGURE 3: ILLUSTRATION OF PRIMARY PROTEIN STRUCTURE FOR A PROTEIN THAT HAS 18 AMINO ACIDS. ..... 7
FIGURE 4: ROC CURVE ILLUSTRATION (REPRODUCED FROM [17]). ..... 11
FIGURE 5: PRECISION RECALL CURVE BASED ON PREVIOUS ROC CURVE (ADAPTED FROM [17]). ..... 12
FIGURE 6: OVERVIEW OF IDEA BEHIND MANY SEQUENCE-BASED PPI PREDICTORS. ..... 16
FIGURE 7: SPACED SEED ILLUSTRATION. ..... 18
FIGURE 8: HIGH LEVEL OVERVIEW OF PIPE. ..... 23
FIGURE 9: ILLUSTRATION OF THE PRECOMPUTATION STEP WITH PIPE FOR A SINGLE PROTEIN. ..... 25
FIGURE 10: OVERVIEW OF PIPE ALGORITHM FOR ONE PAIR OF SLIDING WINDOWS. ..... 26
FIGURE 11 SAMPLE PIPE LANDSCAPES ..... 27
FIGURE 12: TRADITIONAL PIPE SCORE USING FILTERING OF LANDSCAPE ..... 28
FIGURE 13: SW METHOD FOR MODIFYING LANDSCAPE (REPRODUCED FROM [64]). ..... 29
FIGURE 14: SAMPLE ROC CURVE FROM PIPE WITH DIFFERENT SPECIES. ..... 31
FIGURE 15: ORIGINAL PRE-PROCESSING ALGORITHM FOR PIPE. ..... 35
FIGURE 16: ILLUSTRATION OF GENERAL DATABASE FORMAT FOR A PROTEIN ..... 36
FIGURE 17: ORIGINAL PIPE ALGORITHM FOR GENERATING LANDSCAPE. ..... 37
FIGURE 18: MODIFIED DATABASE FORMAT EXAMPLE FOR A SINGLE PROTEIN, TYPE 1 DB FILE. ..... 42
FIGURE 19: MODIFIED PRE-PROCESSING ALGORITHM FOR PIPE TO CREATE TYPE 1 DB FILES. ..... 42
FIGURE 20: MODIFIED DATABASE FORMAT EXAMPLE FOR A SINGLE PROTEIN, TYPE 2 DB FILE ..... 43
FIGURE 21: MODIFIED PRE-PROCESSING ALGORITHM FOR PIPE TO CREATE TYPE 2 DB FILES. ..... 44
FIGURE 22: MODIFIED PIPE LANDSCAPE GENERATION ALGORITHM. ..... 45
FIGURE 23: SW SCORE NORMALIZATION CHANGE FOR CROSS-SPECIES PREDICTION. ..... 59
FIGURE 24: NORMALIZATION FACTOR AS A FUNCTION OF THE NUMBER OF PROTEINS FOR 3 DIFFERENT NUMBERS OF SPECIES. ..... 60
FIGURE 25: EVOLUTIONARY DIVERGENCE TIMELINE ESTIMATES FOR SPECIES USED (GENERATED WITH [69]). ..... 62
FIGURE 26: PREDICTION OF *H. SAPIENS – H. SAPIENS* PPI FROM SEVEN TRAINING SPECIES. ..... 67
FIGURE 27: PREDICTION OF *M. MUSCULUS – M. MUSCULUS* PPI FROM SEVEN TRAINING SPECIES ..... 68
FIGURE 28: ROC CURVE FOR MULTI-SPECIES CROSS-SPECIES PREDICTION OF *H. SAPIENS* PPI SHOWING THE PERFORMANCE OF THE TWO METHODS OF NORMALIZATION. ..... 70
FIGURE 29: ROC CURVES OF USING SEVEN TRAINING SPECIES TO PREDICT INTRA-SPECIES PPI IN THE EIGHTH SPECIES. ..... 70
FIGURE 30: P-R CURVES AT 10:1 CI FOR EACH OF THE TESTING SPECIES WHEN USING THE SEVEN OTHER SPECIES FOR TRAINING DATA. ..... 71
FIGURE 31: P-R CURVE OF USING DIFFERENT SPECIES TO PREDICT KNOWN *H. SAPIENS* INTERACTOME. ..... 73
FIGURE 32: HIGH LEVEL OVERVIEW OF HOW TRAINING DATA WERE USED TO PREDICT INTER-SPECIES PPI BETWEEN SOYBEAN AND SCN USING THE KNOWN INTERACTIONS FROM TWO SIMILAR MODEL SPECIES. ..... 84
FIGURE 33: HIGH-LEVEL OVERVIEW OF THE DATA USED TO PREDICTION INTER-SPECIES PPI BETWEEN SOYBEAN AND HUMAN. ..... 85
FIGURE 34: RECEIVER OPERATING CHARACTERISTIC (ROC) CURVE FOR HIV-1 – HUMAN PPI PREDICTION. 86
FIGURE 35: PRECISION-RECALL CURVE FOR HIV-1 – HUMAN PPI PREDICTION. ..... 87
FIGURE 36: PLOT SHOWING TOTAL NUMBER OF PREDICTED INTERACTIONS AND THE NUMBER THAT ARE EXPECTED TO BE TRUE INTERACTIONS AS A FUNCTION OF RECALL. ..... 88
FIGURE 37: ROC CURVE FOR PREDICTING A. THALIANA FROM EVERY OTHER TRAINING SPECIES. ..... 114
FIGURE 38: ROC CURVE FOR PREDICTING C. ELEGANS FROM EVERY OTHER TRAINING SPECIES. ..... 115
FIGURE 39: ROC CURVE FOR PREDICTING D. MELANOGASTER FROM EVERY OTHER TRAINING SPECIES. .. 116
FIGURE 40: ROC CURVE FOR PREDICTING H. SAPIENS FROM EVERY OTHER TRAINING SPECIES. ..... 117
FIGURE 41: ROC CURVE FOR PREDICTING M. MUSCULUS FROM EVERY OTHER TRAINING SPECIES. ..... 118

FIGURE 42: ROC CURVE FOR PREDICTING P. FALCIPARUM FROM EVERY OTHER TRAINING SPECIES. ..... 119
FIGURE 43: ROC CURVE FOR PREDICTING S. CEREVISIAE FROM EVERY OTHER TRAINING SPECIES. ..... 120
FIGURE 44: ROC CURVE FOR PREDICTING S. POMBE FROM EVERY OTHER TRAINING SPECIES. ..... 121
FIGURE 45: P-R CURVE FOR PREDICTING A. THALIANA FROM EVERY OTHER TRAINING SPECIES (AT 10:1 CI). ..... 122
FIGURE 46: P-R CURVE FOR PREDICTING C. ELEGANS FROM EVERY OTHER TRAINING SPECIES (AT 10:1 CI). ..... 123
FIGURE 47: P-R CURVE FOR PREDICTING *D. MELANOGASTER* FROM EVERY OTHER TRAINING SPECIES (AT 10:1 CI). ..... 124
FIGURE 48: P-R CURVE FOR PREDICTING *H. SAPIENS* FROM EVERY OTHER TRAINING SPECIES (AT 10:1 CI). 125
FIGURE 49: P-R CURVE FOR PREDICTING *M. MUSCULUS* FROM EVERY OTHER TRAINING SPECIES (AT 10:1 CI). ..... 126
FIGURE 50: P-R CURVE FOR PREDICTING *P. FALCIPARUM* FROM EVERY OTHER TRAINING SPECIES (AT 10:1 CI). ..... 127
FIGURE 51: P-R CURVE FOR PREDICTING *S. CEREVISIAE* FROM EVERY OTHER TRAINING SPECIES (AT 10:1 CI). ..... 128
FIGURE 52: P-R CURVE FOR PREDICTING *S. POMBE* FROM EVERY OTHER TRAINING SPECIES (AT 10:1 CI).. 129
FIGURE 53: P-R CURVE OF PREDICTING A. THALIANA INTERACTIONS FROM RANDOM SAMPLE OF 2000 INTERACTIONS FROM EACH TRAINING SPECIES AT 10:1 CI (REPEATED 20 TIMES, CURVES AVERAGED). ..... 131
FIGURE 54: P-R CURVE OF PREDICTING C. ELEGANS INTERACTIONS FROM RANDOM SAMPLE OF 2000 INTERACTIONS FROM EACH TRAINING SPECIES AT 10:1 CI (REPEATED 20 TIMES, CURVES AVERAGED). ..... 132
FIGURE 55: P-R CURVE OF PREDICTING D. MELANOGASTER INTERACTIONS FROM RANDOM SAMPLE OF 2000 INTERACTIONS FROM EACH TRAINING SPECIES AT 10:1 CI (REPEATED 20 TIMES, CURVES AVERAGED). ..... 133
FIGURE 56: P-R CURVE OF PREDICTING H. SAPIENS INTERACTIONS FROM RANDOM SAMPLE OF 2000 INTERACTIONS FROM EACH TRAINING SPECIES AT 10:1 CI (REPEATED 20 TIMES, CURVES AVERAGED). ..... 134
FIGURE 57: P-R CURVE OF PREDICTING M. MUSCULUS INTERACTIONS FROM RANDOM SAMPLE OF 2000 INTERACTIONS FROM EACH TRAINING SPECIES AT 10:1 CI (REPEATED 20 TIMES, CURVES AVERAGED). ..... 135
FIGURE 58: P-R CURVE OF PREDICTING S. CEREVISIAE INTERACTIONS FROM RANDOM SAMPLE OF 2000 INTERACTIONS FROM EACH TRAINING SPECIES AT 10:1 CI (REPEATED 20 TIMES, CURVES AVERAGED). ..... 136
FIGURE 59: P-R CURVE OF PREDICTING S. POMBE INTERACTIONS FROM RANDOM SAMPLE OF 2000 INTERACTIONS FROM EACH TRAINING SPECIES AT 10:1 CI (REPEATED 20 TIMES, CURVES AVERAGED). ..... 137
FIGURE 60: P-R CURVE OF PREDICTING P. FALCIPARUM INTERACTIONS FROM RANDOM SAMPLE OF 2000 INTERACTIONS FROM EACH TRAINING SPECIES AT 10:1 CI (REPEATED 20 TIMES, CURVES AVERAGED). ..... 138

# List of Abbreviations

| Abbreviation | Definition |
|---|---|
| AA | Amino acids |
| AUC-ROC | Area under the ROC curve |
| AU-PRC | Area under the P-R curve |
| CI | Class imbalance |
| FPR | False positive rate (also known as 1 – specificity) |
| HDD | Hard disk drive |
| HIV-1 | Human Immunodeficiency Virus, type 1 |
| LOOCV | Leave-one-out cross-validation |
| PIPE | Protein-protein interaction prediction engine |
| PPI | Protein-protein interaction(s) |
| Pr | Precision |
| P-R curve | Precision-recall curve (precision vs TPR curve) |
| Pr@25%TPR | Precision at 25% TPR (performance metric) |
| RAM | Random access memory |
| ROC curve | Receiver operating characteristic curve (TPR vs FPR curve) |
| SCN | Soybean Cyst Nematode (soybean pest) |
| SSD | Solid-state drive |
| SW score | Similarity-weighted score |
| TPR | True positive rate (also known as sensitivity, recall) |

# 1 Introduction

## 1.1 Introduction

The soybean legume (*Glycine max*) is one of the most valuable crops globally. This high-protein oilseed is used for many purposes, including human consumption, animal feed, and biodiesel [1]. The United States is the largest soybean producer in the world, while Canada ranks as the 7$^{th}$ largest producer, exporting more than 1 billion USD worth of soybeans annually [2]. Additionally, the price of soybeans is projected to rise at a rate of 0.4% annually, whereas the projected prices of other leading Canadians crops such as wheat and corn are expected to decline [3]. The Canadian soybean market is currently concentrated in the province of Ontario, which is responsible for over half of Canada's annual soybean production. However, many Ontario farmers are plagued by the Soybean Cyst Nematode (SCN; *Heterodera glycines* Ichinohe), a detrimental pest capable of reducing yield up to 40% per field [4]. The SCN pest was first identified in Canada in 1998, and has since become widespread across Canadian soybean fields. In the United States and Canada, it causes more than twice as much yield loss as any other disease or pest [5]. In the United States alone, it is estimated to cause more than 1 billion USD in lost soybean yield per year [6]. Figure 1 shows the distribution of the pest across Canada and the United States. The SCN pest has also recently been seen further north in Quebec [7].

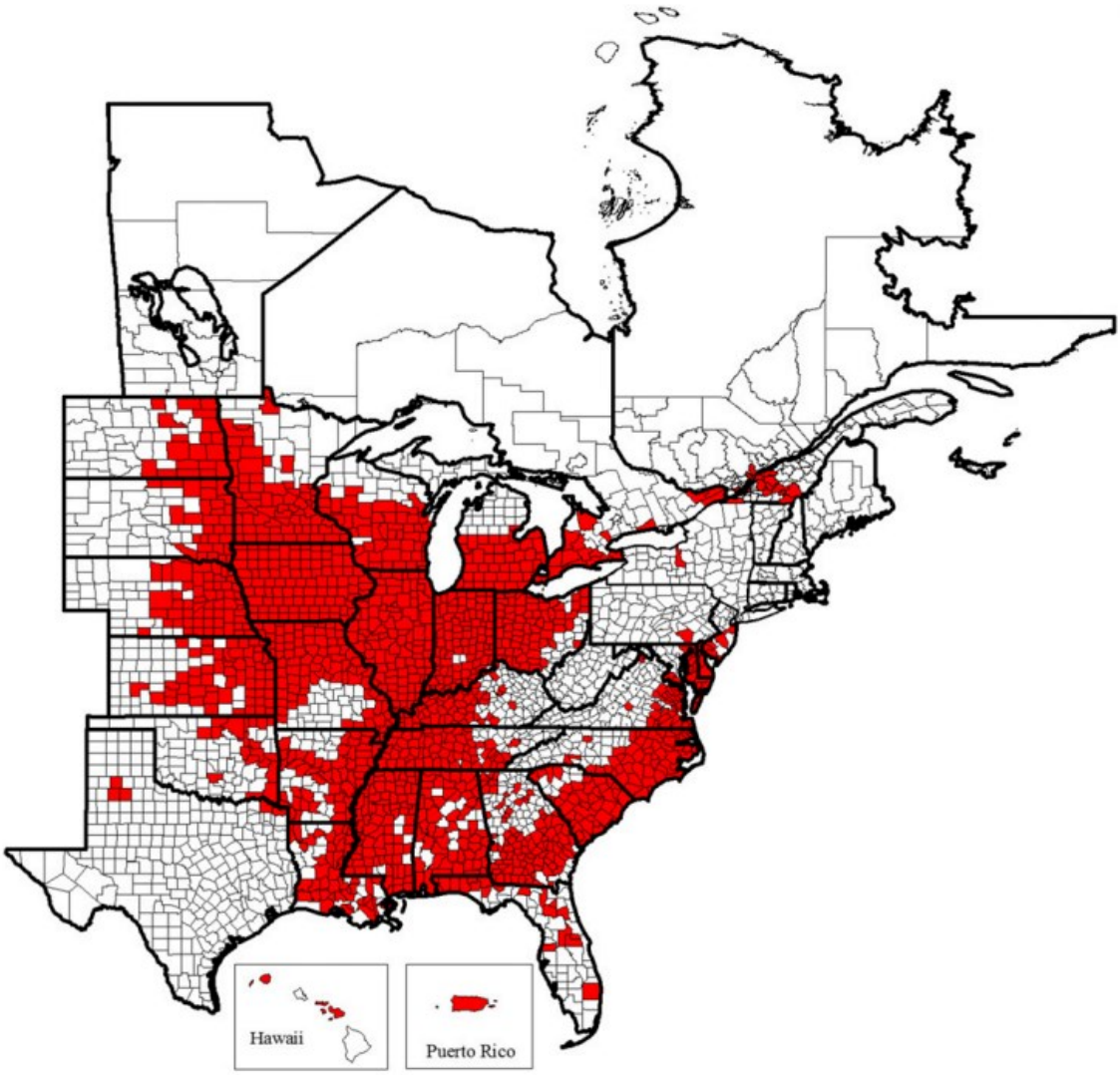



**Figure 1: Prevalence of SCN across counties in Canada and the US (reproduced from** [8]).

As such, there is significant ongoing research into combatting the pest and developing SCN-resistant strains. In Ontario, resistant strains such as the PI88788 and Peking sources exist, however they do not completely prevent yield loss due to SCN [4]. Unfortunately, SCN varieties also exist which are not impeded by these resistant strains. Agriculture and Agri-Food Canada aims to develop SCN-resistant strains and learn more about the specific molecular mechanisms of SCN infection. In 2016, they approached the

Carleton University Bioinformatics Research Group to form a research partnership dedicated to combatting SCN and developing more resistant soybean strains.

Protein-Protein Interactions (PPI) are expected to play a major role in the mechanism of infection in soybean crops by SCN, given that they are known to play a role in many infections by viruses, bacteria, and other plant pests. However, experimentally determining PPI is costly and time-consuming for even a small number of PPI. Given that Soy has approximately 75,000 proteins and SCN has approximately 22,000, there are over 1.6 billion possible interactions between SCN and soybean proteins. Therefore, computational methods must be used to guide and prioritize experiments. Carleton's role in the Agriculture and Agri-Food Canada SCN resistance project is to predict novel PPI between SCN and soybean using leading computational methods. However, most computational methods of predicting PPI were designed and validated strictly within well-studied species using abundant known data. Since soybean and SCN are both relatively under-studied with limited available data, predicting novel PPI is a difficult challenge. This thesis will examine transferring data from other species in order to make novel PPI predictions between SCN and soybean.

## 1.2 Motivation

The Protein-protein interaction prediction engine (PIPE) [9], developed by members of Carleton University Bioinformatics Research Group, is a leading method of PPI prediction method both in terms of speed and accuracy. The PIPE algorithm may be suitable for predicting PPI between SCN and soybean for several reasons. The first is the fast, parallel implementation of PIPE. PIPE is able to predict the entire human interactome (the complete set of all 254 million potential human-human PPI) in under one month using

a reasonably-sized computer cluster [10]. Such speed is necessary for SCN-soybean prediction as there are over 1.6 billion PPI to examine, a significantly larger task than predicting the human interactome. The second reason is PIPE's minimal input data requirements. Only primary sequence and a list of known PPI are required, which makes it possible to make predictions in under-studied species. The third reason is that PIPE has previously shown capable of making predictions in one species using the known interactions from another species (termed **cross-species prediction**) [11]. This ability is necessary to make SCN-soybean PPI predictions, as neither SCN nor soybean have known PPI to use to predict novel PPI.

However, the use of PIPE to predict **inter-species PPI**, where the interacting proteins arise from two different species, has not been investigated. Furthermore, only preliminary investigation of PIPE's ability to predict cross-species PPI has been conducted, concluding solely that PIPE's performance was better than random. The prediction of soybean-SCN PPI involves using cross-species data to predict inter-species PPI, as illustrated in Figure 2. Since there are few to no known interactions between SCN and soybean, experimental validation of PPI predictions is not possible at this time. Therefore, prior to undertaking these predictions, it is necessary to further characterize PIPE's performance in these two tasks, to develop best practices for this type of cross- and inter-species PPI predictions, and possibly to improve PIPE in these regards. This represents the central goals of the present thesis.

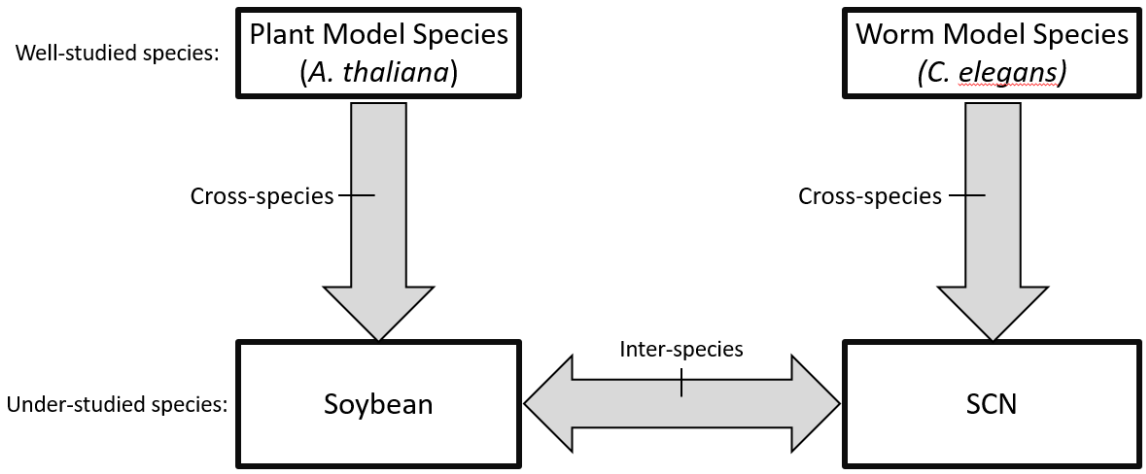



**Figure 2: High-level overview of cross-species and inter-species PPI prediction. Cross-species prediction involves taking data from a well-studied species (such as *A. thaliana* or *C. elegans*) and transferring it to under-studied species such as soybean and SCN. Inter-species prediction is when interactions are predicted between proteins in different species, such as soybean and SCN.**

With the growing threat to the soybean industry from SCN, prediction of all the interactions between soy and SCN is a critical next step in the development of resistant strains. Additionally, if PIPE is shown capable of predicting PPI between under-studied species, then there are many other potential use-cases in which PIPE could be applied, including crop-pest PPI prediction for other crops, human-pathogen PPI prediction, and livestock-pathogen PPI prediction.

## 1.3 Problem Statement

Motivated by the specific problem of soybean-SCN PPI prediction, this thesis seeks to develop best practices for using PIPE to make both cross-species and inter-species PPI predictions. Specifically, we seek to improve PIPE to make it faster and more accurate for such cases.

## 1.4 Organization of Thesis

This thesis consists of six chapters. Chapter 2 provides a brief introduction to protein biology and PPI; a brief overview of general pattern classification as it applies to PPI prediction; a review of different PPI prediction methods leveraging structure and sequence; and an overview of the PIPE algorithm. In Chapter 3, we explain the detailed implementation of the PIPE algorithm, and propose a modification that significantly reduces the time required to run the algorithm. In Chapter 4, we discuss cross-species PPI prediction, which involves predicting one species' complete interactome (the set of all PPI in that species) from the known PPI of another species. We propose a change in the PIPE scoring algorithm specific to cross-species PPI prediction and then examine the accuracy of the PIPE algorithm when using different species of varying evolutionary distance. In Chapter 5, we apply the PIPE algorithm to inter-species PPI prediction, where PPI between different species are predicted. We first examine predicting PPI between the HIV-1 virus and human. We also examine the motivating example of Soy-SCN PPI prediction, which involves both cross-species and inter-species PPI prediction. Chapter 6 presents a summary of contributions and provides recommendations for future work.

# 2 Background and Literature Review

## 2.1 Background Biology

Proteins are large biomolecules which perform a wide variety of tasks in every organism, including catalysing chemical reactions, providing structure for cells, transmitting signals, transporting molecules, regulating cell division, and decomposing old proteins to build new proteins. A protein is built from amino acid residues which are chained together in an order specified by a sequence of DNA known as a gene. The gene contains instructions on which of the twenty amino acids should be linked together in what order. The string of amino acid residues that make up a protein is known as the primary structure of the protein. Each of the twenty different amino acids is commonly represented by a single letter (A, R, N, D, C, E, Q, G, H, I, L, K, M, F, P, S, T, W, Y, V). The twenty different amino acids all have different chemical properties which affect how the protein folds to a final 3D structure, which then affects the role and function of the protein. An illustration of the primary structure of a protein is given in Figure 3.


**Figure 3: Illustration of primary protein structure for a protein that has 18 amino acids. Each blue circle represents a specific amino acid residue indicate by the letter. The amino acid residues are all chained together.**

Proteins are very important for studying an organism; however, the number of proteins in an organism does not always correlate to the expected complexity of the organism [12]. For example, there are approximately 25,000 genes in *H. sapiens*, but approximately 21,000 genes in *C. elegans*, an arguably simpler species of nematode. Rice is expected to contain approximately 50,000 genes, and soybean approximately 75,000

[13]. The size of the interactome – which is the set of all protein-protein interactions (PPI) in an organism – has shown to better correlate with the complexity of an organism [12]. Proteins can bind together to form multi-protein complexes; they can also interact for signal transduction, expression level regulation, and other purposes. Additionally, many viruses and other pathogens affect their host largely through PPI between pathogen and host proteins [14]. Knowledge of the PPI in an organism is therefore very important for understanding how the organism works. However, only a small fraction of the expected interactome is currently known [12].

Unfortunately, experimentally determining PPI is time-consuming and expensive. High throughput methods to determine many novel PPI exist, but their accuracy has been called into question [15]. Methods such as yeast two-hybrid can test whether two proteins interact, but are time-consuming and expensive for even a single pair of proteins [16]. As there are over 200 million possible pairs of proteins in human, experimentally verifying the complete interactome is infeasible. Therefore, computational methods are often used to predict novel PPI in order to guide biological experiments.

## 2.2 Pattern Classification Overview

Pattern classification is an area of machine learning that seeks to make predictions based on patterns in data. One approach to pattern classification is supervised learning, in which one attempts to predict the class of new observations of data based on a set of training data where the classes are already known. The task of predicting novel PPI is an example of supervised learning: specifically, binary classification. The training data consists of a list of protein pairs, feature data describing each pair, and a class label indicating whether the proteins interact or not. From the training data, new pairs can be

classified as interacting or not. The case of predicting an entire interactome is generally done by examining every pair of proteins and predicting whether they interact or not. This is commonly done by generating a score for each pair of proteins. If the score for a pair of proteins is above a threshold, then the pair is said to interact. If the score is below the threshold, the pair is said to not interact. The threshold can be varied depending on the desired performance of the classifier, as detailed below.

The performance of the classifier at different cut-off thresholds can be measured. This done by testing on a set of known positive (physically interacting) and negative (non-interacting) pairs. To ensure independence between the training and testing data, these test pairs should not appear in the training data set. Each of the pairs in the testing set can be categorized as one of the following four cases depending on the true class label and the predicted class label:

- True positive (TP): pair from positive class that was correctly predicted as positive (score above threshold)
- True negative (TN): pair from negative class that was correctly predicted as negative (score below threshold)
- False positive (FP): pair from negative class that was incorrectly predicted as positive (score above threshold; Type I error)
- False negative (FN): pair from positive class that was incorrectly predicted as negative (score below threshold; Type II error)

From these four basic metrics, more advanced metrics can be developed. Let N be the total number of negative pairs in the test data, and let P be the total number of positive pairs in the test data. The true positive rate (TPR) and false positive rate (FPR) are given by the two equations below:

$$TPR = \frac{TP}{P} = Sensitivity = recall \qquad 2.1$$

$$FPR = \frac{FP}{N} = 1 - Specificity \quad (2.2)$$

The TPR gives the percentage of positives pairs that were correctly labelled as positive in the testing set (the TPR is also commonly known as sensitivity or recall). The FPR is the percentage of negative pairs that were incorrectly labelled as positive in the testing set (a similar metric, specificity, is 1 – FPR). The accuracy of the classifier for a cut-off threshold is given by the equation below:

$$accuracy = \frac{TP + TN}{P + N} \quad (2.3)$$

A perfect classifier would predict every positive pair as positive (TPR = 1) and every negative pair as negative (FPR = 0), giving an accuracy of 1. Classifiers are rarely perfect however, and a trade-off must be made between TPR and FPR. This is done by varying the cut-off threshold through all possible values, measuring the TPR and FPR for each cut-off threshold. The performance across all decision thresholds is often plotted on a received operating characteristic (ROC) curve, as illustrated in Figure 4 below.

A perfect classifier would exist in the top left corner of the ROC curve, at a TPR of 1 and FPR of 0, perfectly dividing the data between positive and negative. A conservative classifier which labels everything as a negative would be in the bottom left corner of the ROC curve, at a FPR of 0 and TPR of 0. A permissive classifier which labels everything as a positive would be in the top right corner of the ROC curve, at a TPR of 1 and FPR of 1. The total area under the ROC curve (AUC-ROC) is often used as a single metric which summarises the entire performance of the classifier (with larger area being more accurate). However, when the predictions are made, the classifier must operate at a single point on the ROC curve. The AUC-ROC curve therefore has limited utility in

evaluating performance, as the classifier will generally operate within a very specific region of the ROC curve.

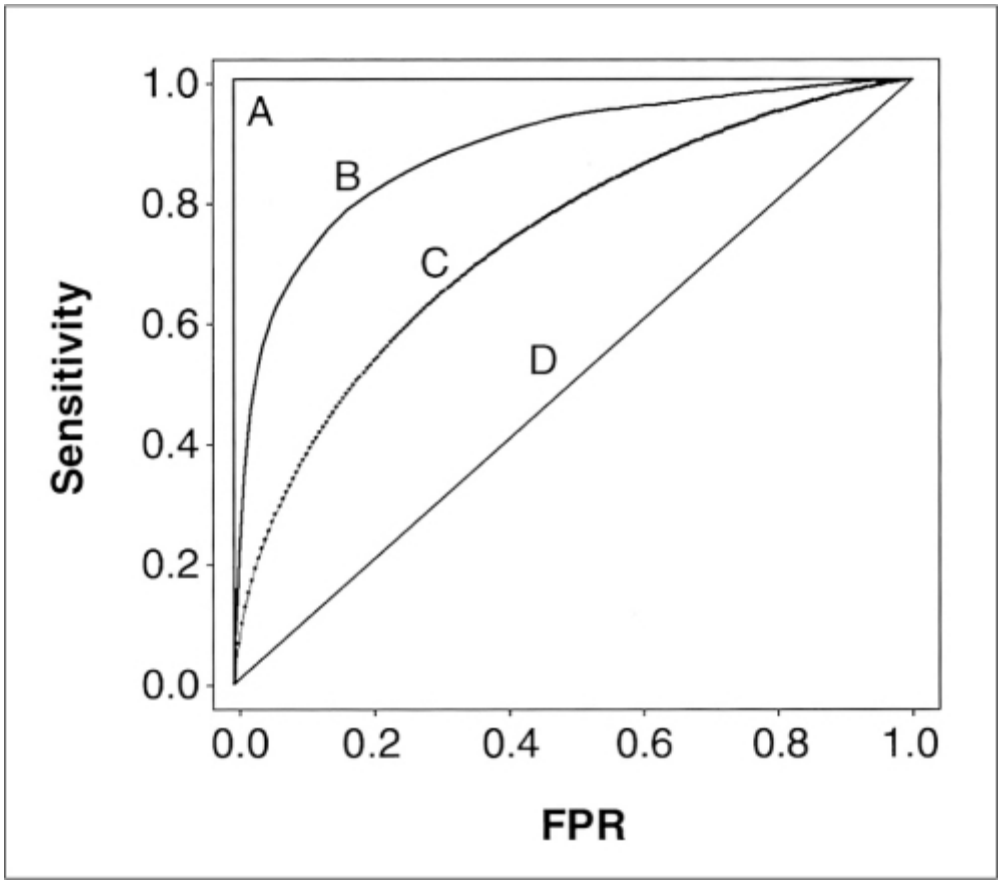



**Figure 4: ROC curve illustration (reproduced from** [17]**). Each line A, B, C and D shows the performance of a different classifier. Line A represents a perfect classifier, where all true interactions are predicted without any false positives. Lines B and C represent two non-perfect classifier; B has higher performance than C at all operating points. Line D represents a classifier made with random guesses.**

In the case of predicting the entire interactome in a species, there are expected to be far more negative pairs that positive pairs. The ratio of negative pairs to positive pairs is known as the degree of class imbalance (CI). In human, it is estimated that there are 300 negative pairs for every positive pair [12]. Therefore, when making novel predictions, the classifier should operate in a range of low FPR, elsewise the number of FP will far outweigh the number of TP when making novel interaction predictions. Both TPR and FPR

fail to reflect the impact of class imbalance. As such, precision (Pr) at the chosen operating point (or cut-off threshold) is generally a preferred metric, as Pr includes class imbalance. The precision is given in the equation below.

$$precision\ (Pr) = \frac{TP}{TP + FP} = \frac{TPR}{TPR + FPR * CI} \quad 2.1$$

In this equation, the CI is given as the number of negatives for every positive. At a given cut-off threshold (and corresponding TPR, FPR, and class imbalance ratio), the precision gives the percentage of predicted positives that are expected to be TP. The precision at different cut-off threshold is commonly displayed in a precision-recall curve (P-R curve or PRC), as seen in Figure 5.

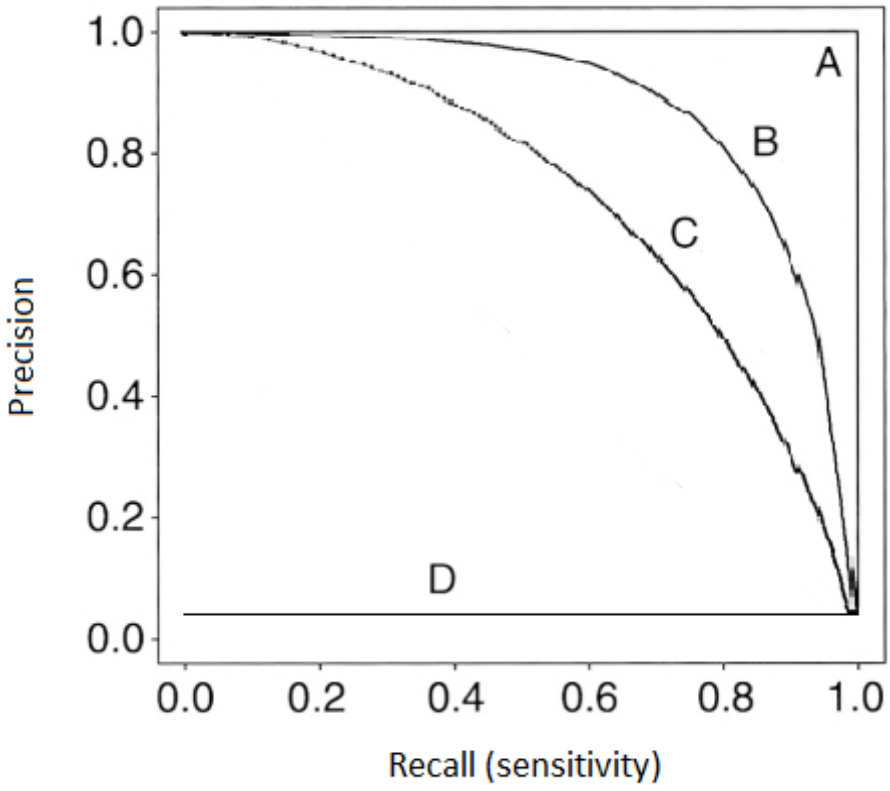



**Figure 5: Precision recall curve based on previous ROC curve (adapted from** [17]**). Line A represents a perfect classifier. Lines B and C represent two non-perfect classifier; B has higher performance that C at all operating points. Line D represents a classifier made with random guesses, for a class imbalance of 20:1.**

This curve shows the percentage of predicted positives that are true positives as a function of the percentage of positives that are correctly predicted as positive. A perfect classifier would be in the upper right corner of the PRC, at a TPR of 1 and precision of 1. A permissive classifier that labels everything as positive would have a TPR of 1, and a precision of CI (which is the percentage of pairs that are expected to truly interact). A random classifier would be a horizontal line at the height of the CI. The area under the PRC (AU-PRC) is a better metric of overall classifier performance

## 2.3 Protein-Protein Interaction Prediction Methods

Protein-protein interaction (PPI) prediction methods can be grouped into two high-level categories: simulation based methods and statistical/machine learning based methods [18]. Simulation based methods rely on molecular simulations of forces between different atoms in the proteins. Due to the high computational cost of such methods, they are rarely used for detecting novel interactions, but rather for estimating binding strength or interaction dynamics [18]. Since protein structures are not known for many proteins, template-based methods exist which seek to use sequence information to find proteins that may have similar structure. One example is PrePPI [19], which makes predictions by using sequence homology to find similar structures for an input pair of proteins. By then comparing the predicted structures with other known structures that are known to interact, the protein pair can be predicted to interact or not. PrePPI claims to be as accurate as experimental high-throughput screening methods [19]. However, template based methods are still limited by the lack of known structures and template structures [18].

The other high-level category is statistical or machine learning based methods. These methods usually rely on a positive set of experimentally verified interactions and a

negative set of protein pairs that are known to not interact in order to make and test novel PPI predictions. Positive interactions are available from several databases, including BioGRID [20], MINT [21], IntAct [22], DIP [23], HPRD [24], and MIPS MPact [25]. These six databases were integrated together by the protein interaction network analysis platform [26]. Specialized databases also exist which include interactions between viruses and hosts [27] and between HIV-1 and human [28]. Finding known protein pairs that do not interact is a more difficult challenge. Due to the class imbalance of the problem, authors commonly take random pairs of proteins that are have not been observed to interact as the negative set [29]. Other authors choose negative pairs from proteins that exist in different cellular locations; however, this may lead to a biased estimate of accuracy [30]. The Negatome database is a manually curated list of proteins pairs that are known to not interact [31].

Methods that rely on known interactions can use a variety of data sources on the proteins to make their predictions, including protein sequence, gene co-expression levels, phylogenetic profiles, evolutionary information, and known binding affinities. These data are commonly used with machine learning techniques, such as support vector machines (SVM), artificial neural networks (ANN), k-nearest neighbours (KNN), and random forests (RF), to make novel PPI predictions (reviewed in [32]). In addition to traditional pattern classification based approaches, other methods leverage sequence interlogs, network topology, and genomic context to make novel PPI predictions. Interlog-based methods rely on finding proteins that are conserved across different species. If two proteins that are conserved across species are known to interact in one species, then they are more likely to interact in the other species as well. In [33] , a protocol was developed to automatically

map interlogs in different species to predict novel PPI. Network topology methods analyze the network of known PPI (where a protein is a node, and a known interaction between proteins is a vertex) in a variety of species to find common features. By comparing these networks to random networks, the methods can assign confidence scores to each protein-protein interaction [34]. From this, novel interactions can be predicted, and interactions can be labelled as false positives. A method which relies on network topology as well as other protein information is explained in [35]. Other PPI prediction methods rely on genomic context. For example, gene neighbouring uses the fact that interacting genes are often located closely together on the genome of a prokaryotic organism to make PPI predictions [36]. Ref [37] describes a method that makes predictions based on gene fusion events. Gene fusion events occur when multiple genes are fused together on the genome, usually indicating a functional association that is conserved evolutionarily. Other genomic context based methods are reviewed in [32].

### 2.3.1 Protein-Protein Interaction Prediction from Sequence

PPI prediction from sequence is a class of widely applicable PPI prediction methods that predict novel PPI from only the primary structure or the sequence of amino acids. Since amino acid sequences are widely available for a variety of species, these methods are widely applicable, able to be used on every pair of proteins with known sequence in a species.

A basic idea behind many sequence-based PPI methods, including [9], [38] is as follows: if two proteins are known to interact, then two proteins that are similar to the known pair are also likely to interact. However, PPI are typically mediated by small sections of each protein that bind together. Therefore, sequence-based methods often

examine domains or small subsequence regions of each protein in order to make predictions. This idea is illustrated in Figure 6, where similar domains are illustrated by like-coloured rectangles. If two proteins, P1 and Q1 are known to interact, then every domain or unique section of protein P1 is more likely to bind with every domain or section of protein Q1. When these pairs of domains are seen in other proteins, those proteins are also more likely to interact. Pairs of domains that might contribute to an interaction are drawn with dotted black lines. By looking at all known PPI, domain pairs that are likely to cause interaction can be found.

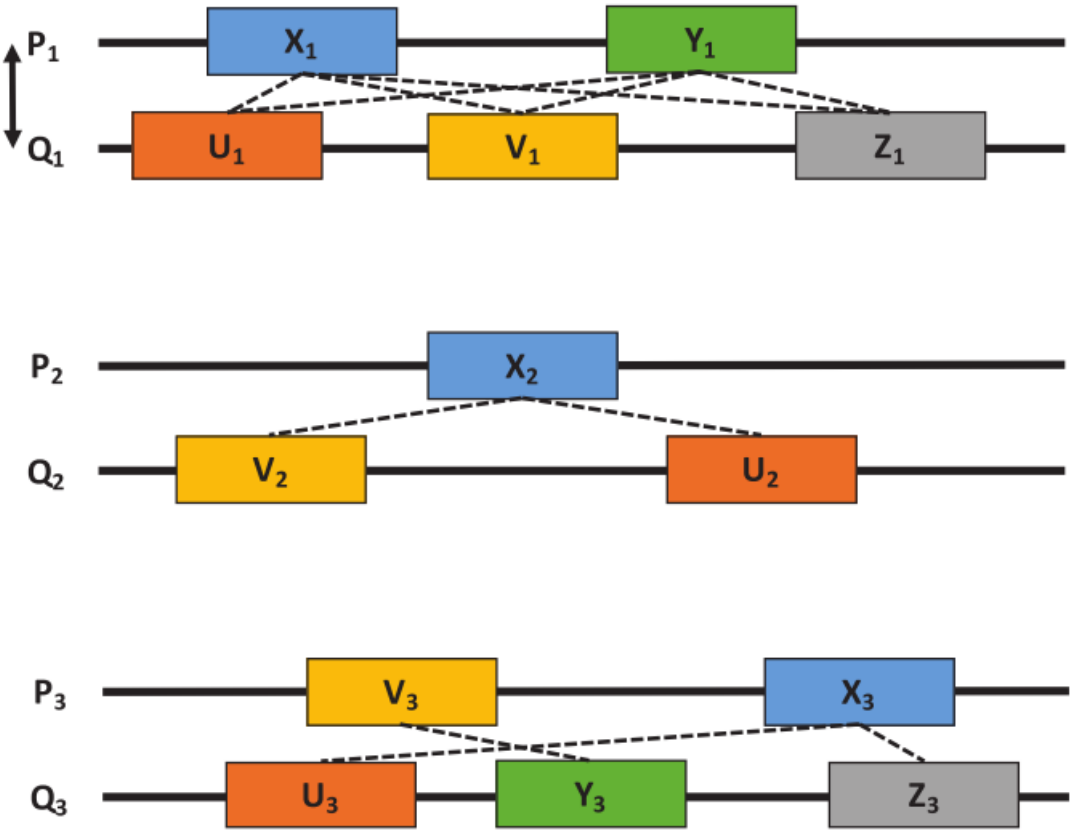



**Figure 6: Overview of idea behind many sequence-based PPI predictors. Proteins P1 and Q1 are known to interact. Similar domains in other proteins are indicated by like-coloured rectangles. When a pair of domains similar to those from an interacting protein is seen in other protein pairs, then that pair is also more likely to interact. (reproduced from** [38]**; image available under Creative Commons Attribution License 4.0: https://creativecommons.org/licenses/by/4.0/)**

Several sequence-based methods are described below; if a method does not have a name, it is referred to by the lead author's name. Martin [39] is an SVM-based method that encodes proteins as vectors without any extra information related to amino acid properties or sequence function. This representation is based off the signature descriptor method which had seen success in the field of chemical informatics. Martin's method was one of the first methods to make a purely sequence-based PPI predictor. Shen [40] is another sequence-based SVM method, similar to Martin. However, instead of using the signature descriptor method, Shen collects amino acids into seven groups based on their physiochemical properties. Shen's method then counts the number of times each possible 3-amino-acid group occurs in a protein, with a total of 7x7x7=343 possible values, and encodes the protein as a frequency vector of the counts. Guo [41] is another SVM-based method similar to Martin and Shen, which encodes proteins using 7 different physiochemical properties. Guo then uses auto-covariance of the sequence with itself at a 30 amino acid offset to generate features for the SVM classifier. These methods, along with the PIPE method (which will be explained in detail in Section 2.3.3) were reviewed and compared by Park in [42]. The PIPE method was shown to have the highest performance among all these methods. Since the Park review, more methods have come been published, including Ding [43], which uses a random forest classifier along with a protein encoding similar to Guo *et al*.

The SPRINT (Scoring PRotein INTeractions) method of predicting PPI is another recent sequence-based method [38], which detects similar subsequences in proteins by using "spaced seeds" [44] to determine sequence similarity. Spaced seeds are illustrated in Figure 7. Only the amino acids at certain locations are relevant (indicated by the non-*

positions within the spaced seed). If the amino acids at the spaced seed points are sufficiently similar (as defined by the PAM120 matrix [45]), then the regions encompassing the space seed are deemed similar. The spaced seeds used by SPRINT were generated automatically for the set of protein sequences examined using the method [46], and are given by {11****11***1, 1**1*1***1*1, 11**1***1**1, 1*1******111}. After similar regions between all pairs of proteins have been identified, SPRINT generates a score for a putative interaction by looping through known interactions and increasing the score of proteins that have similar areas to the interacting pair. After the score has been generated, all pairs that have a score higher than a tunable decision threshold are predicted.

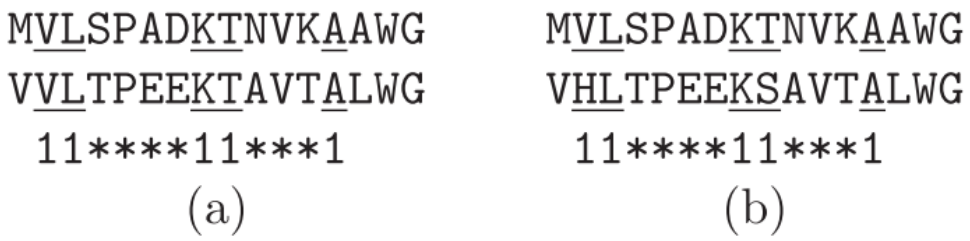



**Figure 7: Spaced seed illustration. (a) shows exact match of two sequences using a spaced seed; (b) shows non-exact but similar match. (reproduced from** [38]**; image available under Creative Commons Attribution License 4.0: https://creativecommons.org/licenses/by/4.0/).**

The SPRINT method [38] compared itself to the previous methods reviewed by Park (which includes an old version of PIPE) as well as the more recent Ding [43] method. SPRINT was shown to be the most accurate on the test cases defined by [29]. Additionally, it was the fastest method listed, able to predict an entire interactome faster than any other method included in the comparison.

### 2.3.2 Intra-, Inter-, and Cross-Species PPI Prediction

An **intra-species** protein-protein interaction is a PPI that exists between proteins in a single species. The prediction of intra-species PPI is the most common application of PPI prediction methods.

Conversely, **inter-species** PPI exist between proteins expressed in different species. For example, the Human Immunodeficiency Virus-1 (HIV-1) infects human cells largely by having HIV-1 proteins interact with human proteins in order to modify host cell functions to express the virus [14]. Knowledge of the PPI network between viruses and their hosts is therefore important for understanding the mechanisms of infection and for the identification of new therapeutic targets [14]. Host-pathogen PPI are the most well-studied example of inter-species PPI, but other cases are possible, such as between symbiotic organisms or as part of the allergic reaction response in human.

**Cross-species** PPI prediction leverages known intra-species interactions from one species to predict intra-species interactions in another species. A common example would be taking interactions from a well-studied model organism to make predictions for an under-studied species.

In a well-studied virus with many known interactions, such as the Human Immunodeficiency Virus-1 (HIV-1), classic machine learning methods have been widely applied to host-pathogen inter-species PPI prediction with limited success (reviewed in [47]). Tastan *et al*. [48] were the first to perform a global PPI prediction between every pair of HIV-1 and human proteins. The authors trained a Random Forest classifier on various types of data, including Gene Ontology (GO) process, expression levels, sequence motifs and the human interactome, and found that they were able to predict PPI with a

Mean Average Precision (MAP) of 23% at a class imbalance of 100:1 (100:1 indicates there are 100 negatives for every positive pair). This work was extended in [49] by having HIV experts manually classify interactions into true interactions and probable interactions based on the amount of published evidence available for each interaction. Using a multilayer perceptron, the MAP was increased to 27.7%. Evans *et al*. [50] predicted interactions using virus and host sequence motifs; a calculated p-value showed predicted interactions affected cellular pathways known to be altered by HIV-1. Dyer *et al.* [51] used an SVM classifier trained with protein sequence k-mers, interaction domain profiles, and the degree of interaction of proteins within the human interactome. This method performed well, finding a precision of 70% at a recall of 40% (also at a class imbalance of 100:1). Mei [52] used probability-weighted ensemble transfer learning to predict interactions; the classifier achieved high recall and specificity, but the data set used for training and testing was artificially class balanced, making the classifier less suitable for global PPI prediction. Nouretdinov *et al.* [53] used conformal prediction to generate a confidence measure for each predicted interaction. Mukhopadhyay *et al.* [54] used an approach based on bi-clustering and rule mining. Structure-based methods such as Doolittle *et al.*[55] and Zhao *et al.* [56] exist, but they are limited to the subset of human and HIV-1 proteins with known structures.

Inter-species PPI prediction has also been applied to a wide variety of different species pairs, including Hepatitis C and human [57], Dengue Fever Virus and human [58], *Arabidopsis thaliana* and the pest *Pseudomonas syringae* [59], and more. DeNovo [60] is a SVM-based method designed to predict interactions between under-studied viruses and

human. DeNovo found that training with interactions between human and viruses similar to the under-studied virus produced the best results.

Cross-species PPI prediction is often done using interlog-based methods. These transfer interactions known to occur in one species by finding similar protein pairs in a new species. Interlog methods usually have a low sensitivity (TPR), but can have a higher precision. Yang *et al*. reviews several interlog-based methods and implement their own Random Forest based method [61]. Chen *et al*. implement a domain-based method for cross-species PPI prediction [62]. The previously mentioned Martin *et al*. method is an SVM-based method that predicts cross-species interactions amongst human, mouse, yeast and *Helicobacter pylori* [39]. Xia *et al*. is a meta-method that uses six other PPI predictors to make cross-species PPI predictions using *Saccharomyces cerevisiae* data for training to predict PPI within *Helicobacter pylori*.

### 2.3.3 The Protein-Protein Interaction Prediction Engine (PIPE)

The Protein-protein Interaction Prediction Engine (PIPE), developed by researchers in the Carleton University Bioinformatics Research Group, is a method for predicting novel protein-protein interactions (PPI). Originally published in [9] as a method for predicting PPI in the yeast species *S. cerevisiae* (more commonly known as Baker's yeast), the computational and classification performance of PIPE have since been improved in [63], [64]. In [65], the method was shown to be applicable to a variety of species including *C. elegans*, *E. coli*, *H. sapiens*, *S. cerevisiae*, and *S. pombe*. In [10], a parallel implementation of the method was used to complete a global prediction of PPI for every possible pair of proteins in *H. sapiens*, the first study of its kind. This effort required over three months of computational time on a SUN UltraSparc T2+ based cluster with 50 nodes, 800 processor

cores and 6,400 hardware-supported threads. The work in this thesis was done in collaboration with the Carleton University Bioinformatics Research Group, and the latest version of the source code for PIPE was received in 2016 from Dr. Andrew Schoenrock, one of the main contributors to PIPE.

PIPE is a sequence-based method for predicting PPI. The only required data are a list of protein pairs that were experimentally determined to interact and the primary structure (or amino acid sequence) of each protein. With these training data, PIPE can produce a score for a given pair of proteins. This score can be used in two main ways: the first would be to rank protein pairs by score, with a higher score indicating a higher chance for the pair to interact. For example, if a researcher knew that a specific protein interacted with an unspecified protein, then a ranked list of pairs could provide guidance about which protein pairs to examine experimentally first. The second way to use the score would be in a binary decision: if the score is higher than a threshold (which varies with use-case), then the pair of proteins is predicted to interact; if it is lower, then they are predicted to not interact. The score cut-off threshold is generally set by using leave-one-out cross-validation (LOOCV) experiments (detailed below) to examine the sensitivity (TPR), specificity (1-FPR), and precision at different score thresholds. This process is shown in Figure 8 below.

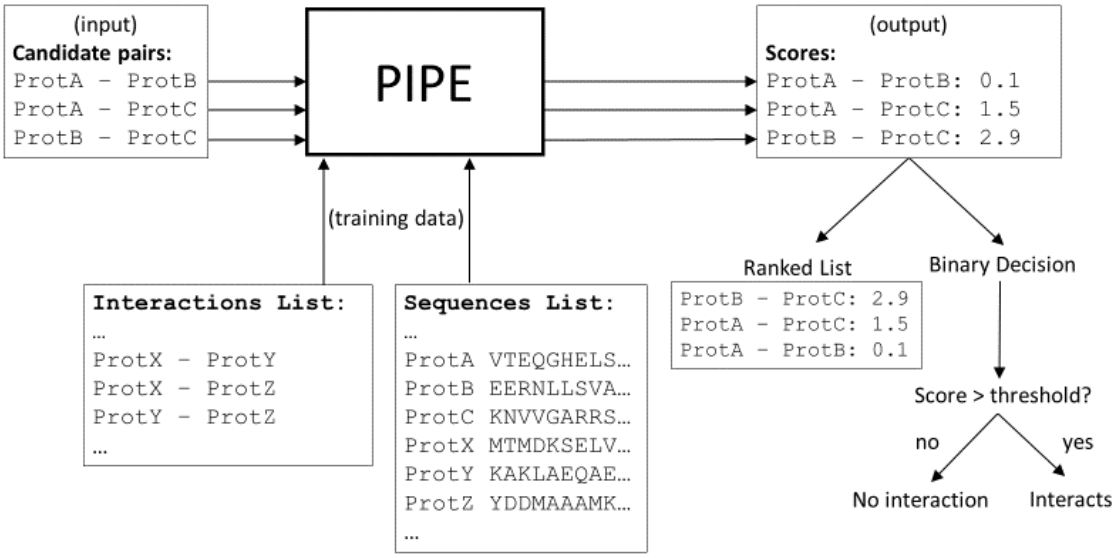



**Figure 8: High level overview of PIPE. PIPE produces a score for a given set of protein pairs using a list of known interactions and the amino acid sequences of each protein. The score can be used to rank the pairs by likelihood of interaction or in a binary decision on whether the pair interacts or not.**

The most common use case for PIPE is a global or all-to-all prediction where PIPE is run on every possible pair of proteins in a species in order to predict the entire interactome, which is the set of all true PPI that occur inside a species. A cut-off threshold is set through LOOCV and then only those pairs with a score above that threshold are predicted to interact. The PIPE algorithm is highly effective for such all-to-all protein interaction predictions due to three features: its high specificity, its minimum input data requirements, and its fast, parallel implementation. Since there are many more proteins that are expected to not interact than to interact, the method must have a high specificity or else the number of false positives would overwhelm the number of true positives, making the all-to-all calculation useless. Many methods rely on 3D protein structure, which is unknown for a large number of proteins. In order to perform a species-wide test, only readily available information, such as amino acid sequence, should be used. Finally, as

there are several hundred million protein pairs to analyze for a typical species, the algorithm must efficient and implemented in parallel to complete in a reasonable time.

The main idea behind PIPE is that interactions are mediated by short subsequences (or "windows") of amino acids, generally where amino acids on the surfaces of the two proteins bind together or otherwise support the interaction. If two proteins are known to interact, then any given subsequence in each of the two proteins is considered to possibly form a complimentary surface or to support the interaction. If these windows are seen in other proteins, then those proteins are also considered more to interact. An alternative formulation of this idea is such: if a pair of windows in a protein pair is often seen in pairs of proteins that are known to interact, then the proteins that have those windows are more likely to interact themselves.

PIPE implements this idea by examining a sliding window of 20 amino acids in each protein, and comparing it to a sliding window of 20 amino acids in every other protein. In a precomputation step, sequence similarity (as measured using the PAM120 matrix [45] with a threshold) is tested between all pairs of sliding windows in all proteins. The results of this precomputation are stored in an efficient binary file for each protein. This way, for each sliding window in a protein, all the proteins which contain a similar window anywhere inside them can be instantly looked up. This process is illustrated in Figure 9 with a sliding window of size 10 amino acids.

In order to generate an interaction score for an input pair of proteins, denoted Protein A and Protein B, every pair of sliding windows between the proteins is examined. Here, $WA_i$ represents the window beginning at position $i$ within protein A. For each pair of sliding windows in the input pair, all proteins that contain a similar window are identified

from the pre-computed list of similar windows. Here, *simprots*($WA_i$) represents the set of proteins containing a window with sequence similarity to $WA_i$. Then, the number of known interactions between all possible pairs from *simprots*($WA_i$) and *simprots*($WB_j$) are counted, and the number is stored in an interaction matrix at location ($i$,$j$).

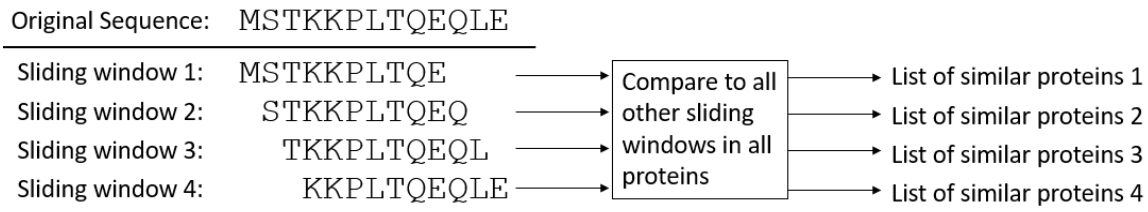



**Figure 9: Illustration of the precomputation step with PIPE for a single protein. There are 4 sliding windows of size 10 amino acids. Each sliding window is compared to all other sliding windows in all proteins. Proteins which contain a sliding window that is similar to the sliding window being examined are stored in a binary database file.**

The PIPE algorithm is illustrated for one pair of windows in Figure 10. The pair of windows being considered is the first 10 amino acids in each protein (coloured in green in the figure). Note that the actual PIPE algorithm uses sliding windows of 20 amino acids; windows of size 10 are used here for clarity). Using the pre-processed similarity database, we can look up and find that proteins *simprots*($WA_1$) = {p1, p5, p9, p12, p24} contain a window that is similar to window $WA_1$ of Protein A. Similarly, the set of proteins *simprots*($WB_1$) = {p2, p4, p9, p22, p64} contain a window similar to $WB_1$ of Protein B. Looking in the known interactions database, we find that there are three interactions between these sets of proteins (p1-p2, p5-p9, and p12-p64). The landscape value at location (1,1) is therefore populated with a 3 for this pair of windows.

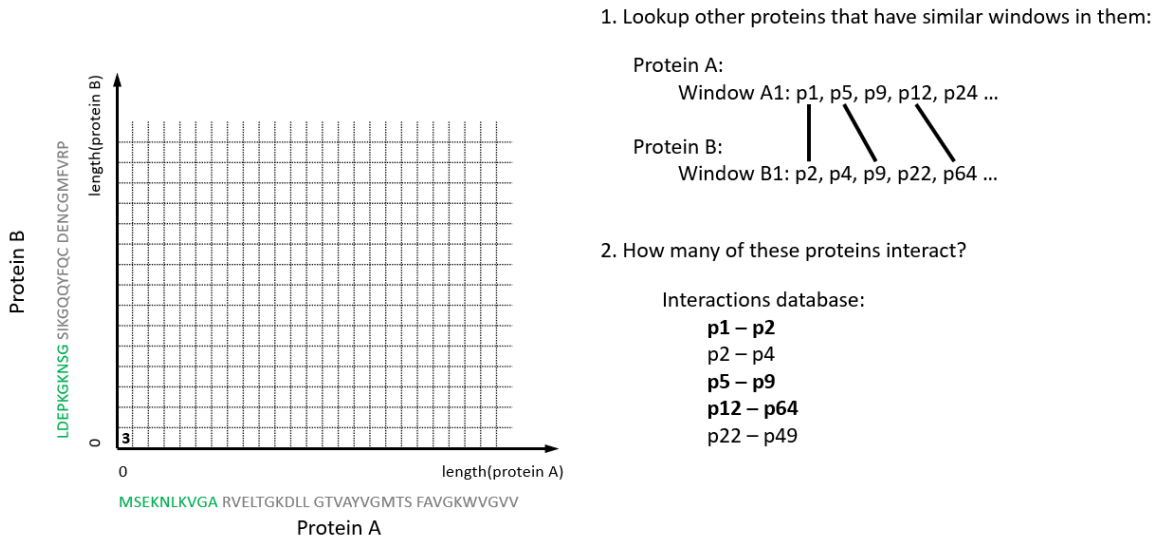



**Figure 10: Overview of PIPE algorithm for one pair of sliding windows. This figure examines the first sliding window of 10 amino acids in each protein, highlighted in green. In step 1, the similar proteins for the windows highlighted in green are identified. In step 2, the known interactions between the similar proteins are counted (the known interactions are highlighted in bold and drawn as black lines between proteins in the two sets). This count is entered in the interaction landscape for these windows. This process is then repeated for each pair of windows.**

The process in Figure 10 in repeated for every pair of windows in the two proteins until a full interaction matrix (also referred to as interaction landscape) is generated. Sample interaction landscapes for novel PPI found with PIPE are shown in Figure 11 below. The landscapes here are illustrated in 3D form, where height on the z-axis is equal the number of 'hits' or known interactions between sets of similar proteins. The height of a peak at location $(i,j)$ represents to weight of evidence that subsequences $WA_i$ and $WB_j$ support a physical interaction between proteins A and B. As discussed below, by aggregating the height of all locations in the landscape, we arrive at a total score for the putative interaction between proteins A-B. Note that it is also possible to estimate the site of interaction within each protein by examining the landscape itself. This method is called PIPE-Sites [66] and is beyond the scope of this thesis.

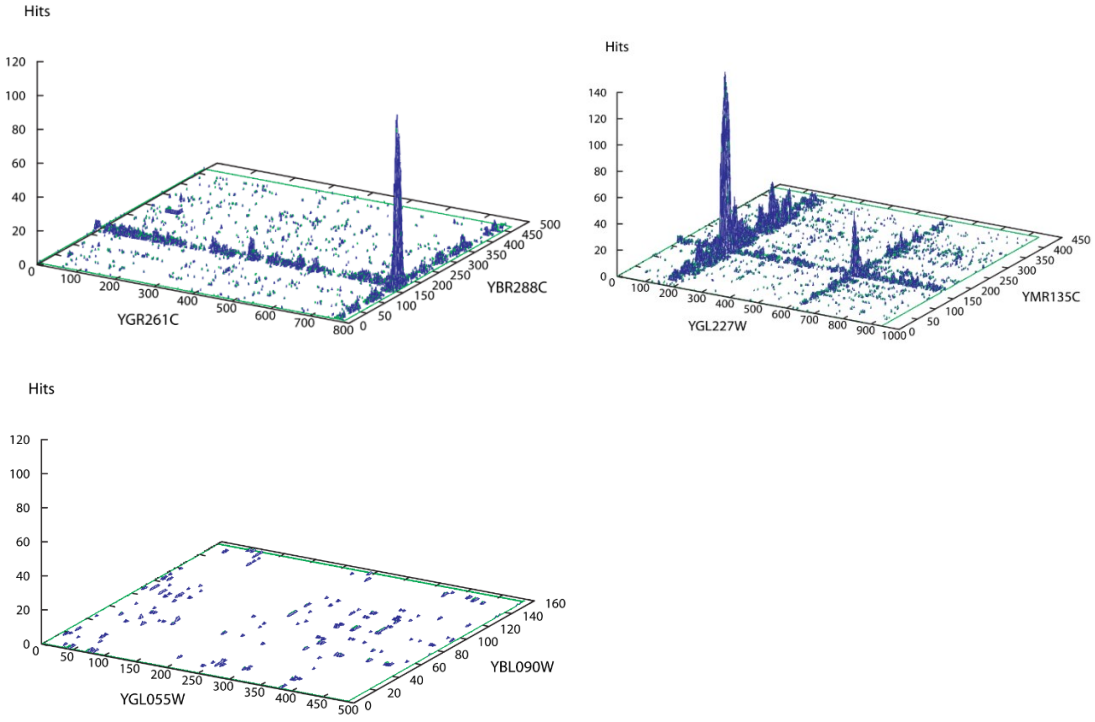



**Figure 11 Sample PIPE landscapes for two positive pairs (top row) and a negative pair (bottom row) (reproduced from** [9]**; images available under Creative Commons Attribution License 2.0: http://creativecommons.org/licenses/by/2.0).**

After generating the full interaction landscape for a protein pair, the landscape is summarized into a single value for ranking or binary classification purposes. This summarization is done with two different methods. The first method, which will be termed the 'traditional PIPE score', involves applying a type of filter to the landscape. The filter is illustrated in Figure 12. It works by setting a matrix element to a value of one if the element and its eight neighbouring entries are all non-zero; otherwise it sets the value to zero. The score is then the average value of the modified landscape.

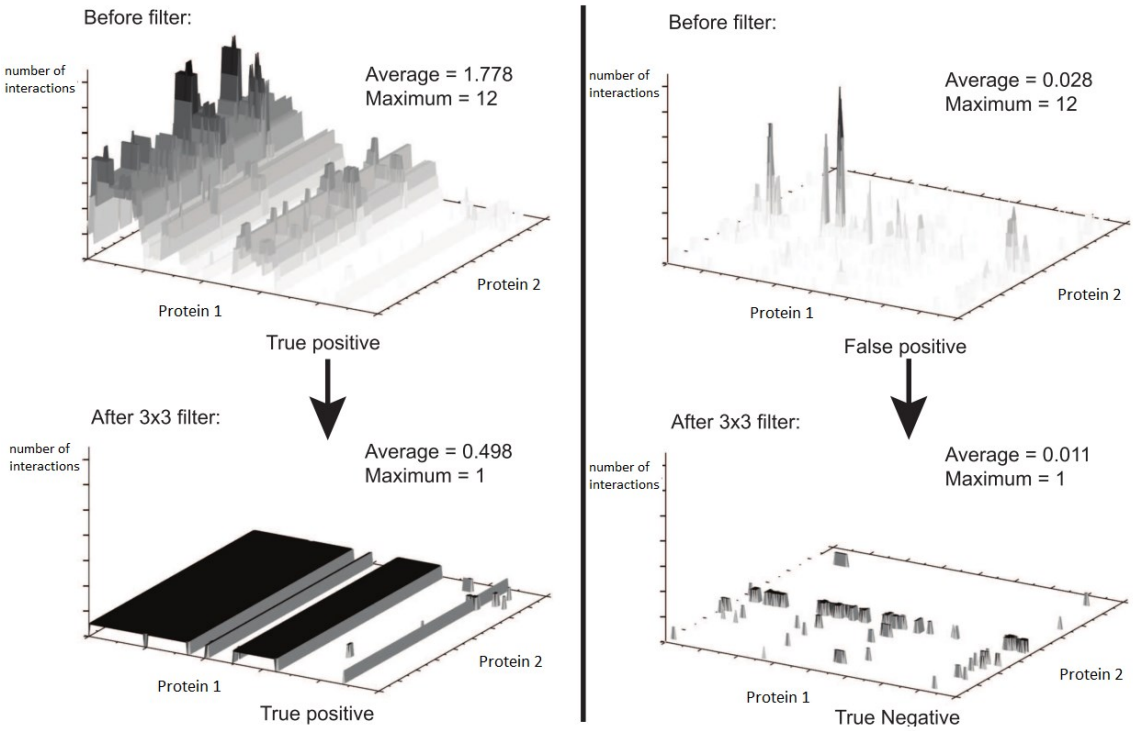



**Figure 12: Traditional PIPE score using filtering of landscape (adapted from** [63]**). The original landscapes are shown in the top row, and both are classified as positives. The bottom row shows the result of applying filter to each of the original landscapes. Now, the right landscape is correctly labelled as a negative.**

The second method of summarizing the landscape into a single score is known as the similarity-weighted (SW) method [64]. This method was developed because certain sliding windows are seen in a very large number of proteins, but are not responsible for supporting or mediating interactions. For example, common subsequences may have other non-PPI functions, such as sub-cellular localization signalling. The SW method normalizes the height of the landscape at a point by counting how many possible interactions there would be if every pair of similar proteins interacted. For location $(i,j)$ this normalization factor would be $|simprots(WA_i)|*|simprots(WB_j)|$. This removes the effect of windows that are very common but do not cause interactions, and strengthens the effect of windows that are relatively rare, but are frequently seen in known interactions. This method is illustrated for a single pair of windows in Figure 13. After the score for each pair of windows is

divided by the normalization factor, the overall SW score is simply the average value of the landscape. The SW method was developed in [64] and compared to the original method; overall the SW method was shown to be superior the original method at all operating points, and is the main score used by PIPE.

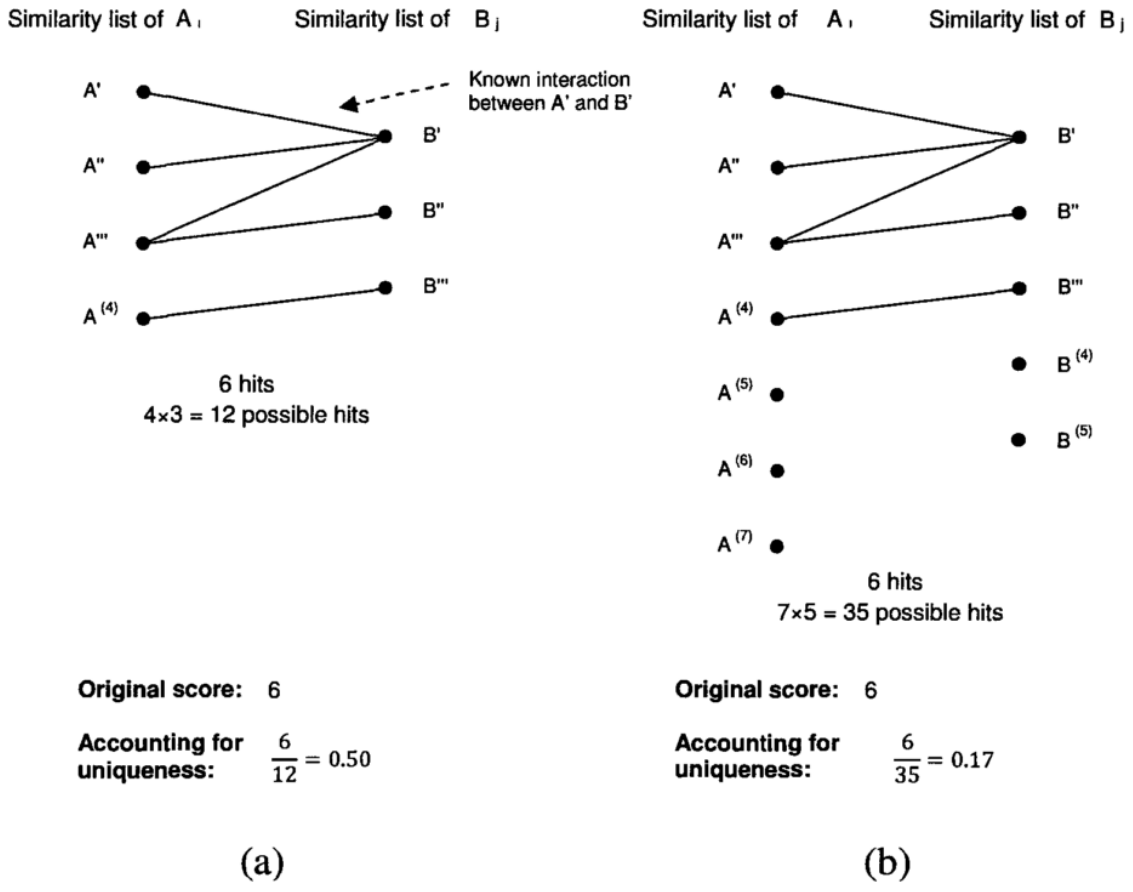



**Figure 13: SW method for modifying landscape (reproduced from** [64]**). The similarity list of $A_i$ given in figure is referred to as *simprots*($WA_i$) in this work.**

As briefly mentioned above, in order to set a score threshold for classification of protein pairs as interacting or not, leave-one-out cross-validation (LOOCV) is used. LOOCV provides the best estimate of future performance, as it simulates how PIPE will

perform when predicting a single new interaction that is not in the training data. Due the implementation of PIPE, LOOCV is very fast to perform. This procedure involves generating the PIPE score for each pair of proteins that are known to interact (the training data). When generating the score for a positive pair, that pair is held out of the training data. After finding the PIPE score for every pair of proteins that are known to interact, the PIPE score is found for a set of negative pairs that are believed to not interact. Since there are very few pairs that have been experimentally determined to not interact, random protein pairs are used as the negative set. Since the class imbalance of the problem is very high, a randomly selected protein pair is very unlikely to be a true interaction. The positive and negative protein pairs are then sorted in descending order of PIPE score. By moving the score threshold from the maximum score to the minimum score, the sensitivity (or TPR) and specificity (1-FPR) can be found for every threshold, forming an ROC curve. Figure 14 shows the ROC curve for different species previously tested with PIPE. With an estimate of the class imbalance of the problem, a P-R curve can also be created. Using these curves, a score cut-off threshold can be selected based on the desired performance. For a given cut-off threshold, the most important metrics are the percentage of true PPI that were successfully predicted by PIPE (sensitivity), and the percentage of PPI predicted by PIPE that are expected to be true PPI (precision).

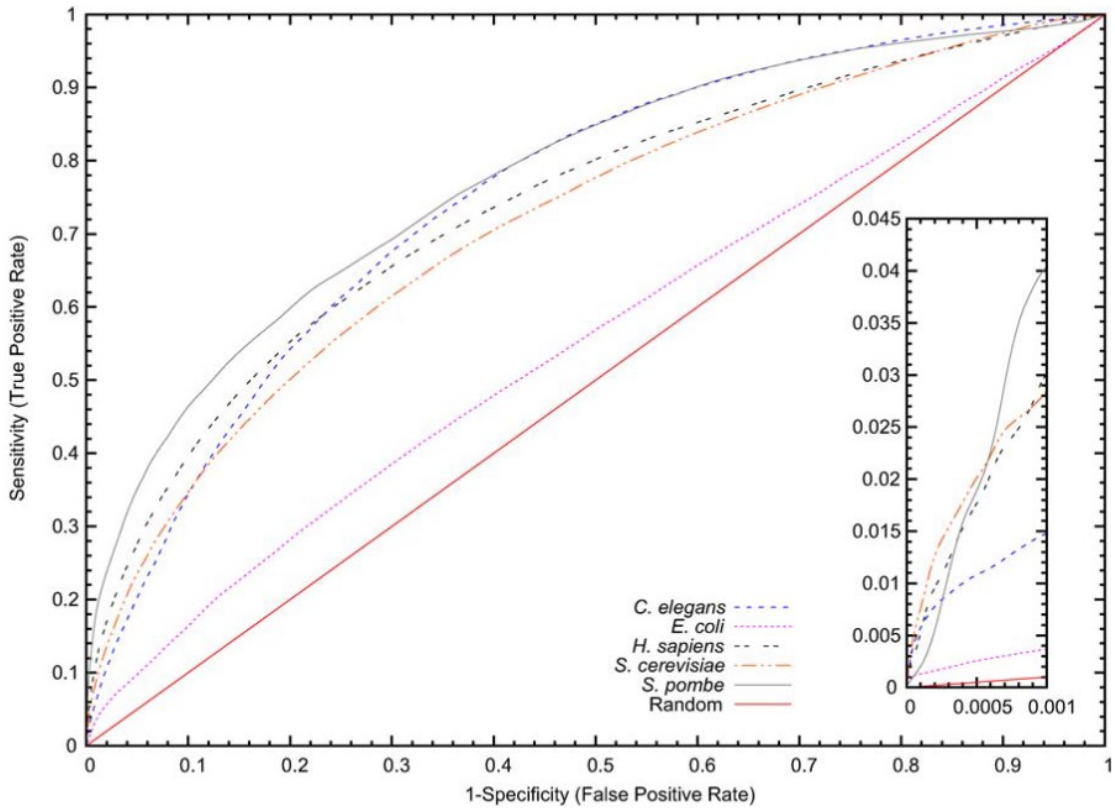



**Figure 14: Sample ROC curve from PIPE with different species. The box in the bottom right shows the area the threshold is generally set in (reproduced from** [65]**).**

### 2.3.4 Summary of the State of the Art

This chapter provided an overview of pattern classification and covers several sequence-based PPI predictors. At the time of this research, PIPE was a leading sequence-based PPI predictor, faster and more accurate than other methods. However there are still limitations with PIPE. While PIPE is parallelized and has been optimized in the past, prediction runs are still excessive for large runs, such as the prediction between SCN – soybean or a human all-to-all (taking several weeks with even hundreds of parallel processes). While PIPE has been used for cross-species prediction in the past, the results were very preliminary and no attempts were made to optimize PIPE for this application.

Additionally, PIPE has never before been used for inter-species prediction. These topics will form the basis for the remainder of this thesis and the contributions therein.

# 3 Computational Improvement of PIPE

## 3.1 Introduction

Over the years, PIPE has been applied to larger and more complex PPI prediction problems. Originally, it was used to predict a small number of pairs within a single species [9]. In 2008, it was used to predict the entire *S. Cerevisiae* interactome [63]. To explore an entire interactome in a given species requires examining a total of $\frac{(N^2+N)}{2}$ pairs, where N is the number of proteins in the species (this equation is essentially $\binom{N}{2}$ pairs, plus N pairs corresponding to a protein interacting with itself). As yeast has approximately 6,721 proteins, the total number of pairs to examine was 22,589,281. After a parallel implementation of PIPE was created, the entire *H. sapiens* interactome was predicted (as N=20,238, the total number of pairs was 204,798,441) [67]. This effort required over three months of computational time on a SUN UltraSparc T2+ based cluster with 50 nodes, 800 processor cores and 6,400 hardware supported threads. Since then, the computational performance of PIPE has been further improved. However, PIPE is also expected to handle even larger interactomes, such as the entire soybean (*G. max*) interactome (with N=75,781, the total number of pairs is 2,871,417,871 – over 10 times larger than the *H. sapiens* interactome). The inter-species predictions desired by Agriculture and Agri-Food Canada researchers are between pairs species with large proteomes such as soybean – human (*G. max – H. sapiens*) and soybean – SCN (*G. max – H. glycines*), which have a similar number of pairs to the soybean interactome. These inter-species prediction runs with PIPE would take several months on a medium sized compute cluster (e.g. 20 8-core CPUs). Furthermore, since both soybean and SCN are being actively investigated by several

research groups, their proteomes and sets of known interactions continue to evolve every few months. This requires periodic rerunning of the entire PIPE run to leverage these new data. As such, the PIPE algorithm is analyzed here and improved in order to complete these inter-species predictions in reasonable timeframe.

In Section 3.2, the current PIPE implementation will be explained (termed "original PIPE"). In Section 3.3, a modified implementation of PIPE with a different core data structure will be proposed. In Section 3.4, the benchmarking methodology for the two implementations will be explained. In Section 3.5, a comparison of the two implementations will be given. In Section 3.6, these chapter will be concluded.

## 3.2 PIPE Algorithm

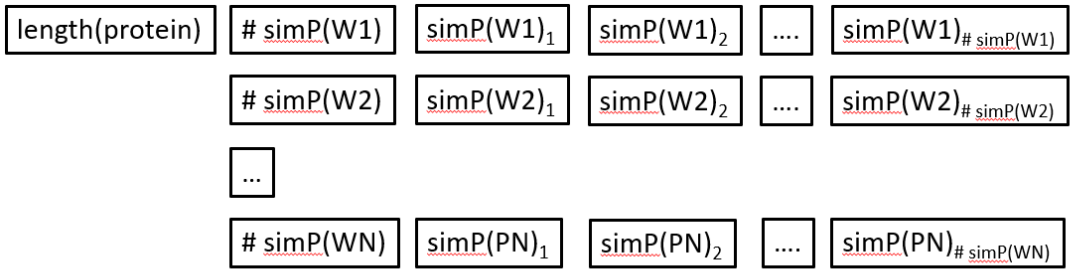
### 3.2.1 Preprocessing and Input Data Representation

Before PIPE can be run on protein pairs to predict interactions, a pre-processing step must occur. This pre-processing step involves calculating which proteins contain similar windows of amino acids (as defined by the PAM120 amino acid substitution matrix and a threshold). This algorithm is show in Figure 15. It creates a database file for each protein; the database file contains a list of similar proteins for each sliding window of amino acids within the protein. A protein is added to the list of similar proteins if it contains a window at any offset that is similar to the specific window of the database protein.

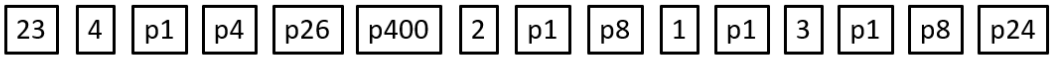
```
     Input
       list of proteins and sequences
     Output
       database file for each protein listing similar proteins for each
     sliding window

1    for each protein, p1 in list of proteins:
2      for each sliding window of amino acids, W1x in p1:
3        let similarProts = [] // empty list of proteins
4        for each protein, p2 in list of proteins:
5          let p2hasAnySimilar = False
6          for each sliding window of amino acids, W2y in p2:
7            if similarity(W1x,W2y) > threshold:
8               set p2hasAnySimilar = True
8          if p2hasAnySimilar == True:
9             add p2 to similarProts
10       write to db file for p1 with len(similarProts), similarProts
         // this contains list of similar proteins for window p1x of p1
```

**Figure 15: Original pre-processing algorithm for PIPE. This algorithm creates a database file for each protein that contains a list of similar proteins for each sliding window of the database protein.**

The database file is written as a ragged array of unsigned integers in a binary file format. The general database format for a single protein is shown in Figure 16 (a) below. The length of the protein is written to the database file, followed by the number of similar proteins (#simP) to the first window (W1), followed by the indices of these similar proteins and so on. Figure 16 (b) shows a toy example of the database file for a protein. This invented protein is 23 amino acids long; however, since the sliding window size used in PIPE is of length 20 amino acids, there are only 4 sliding windows included (AA1-AA20, AA2-AA21, AA3-AA22, AA4-AA23). The first sliding window contains 4 similar proteins, proteins p1, p4, p26, p400. The 'p' prefix is added for clarity. These proteins are written to file as a single unsigned integer based on their index in the protein sequences file. The second window contains 2 similar proteins, p1 and p8. The third sliding window contains 1 similar protein, protein p1. The final sliding window contains 3 similar proteins, proteins p1, p8 and p24.




(a)




(b)

**Figure 16: Illustration of general database format for a protein (a). For each sliding window, similar proteins (simP) are stored. (b) Toy example of database format for a protein with length 23 amino acids. Each black box represents an unsigned integer written to the binary database file.**

In order to run PIPE on a pair of proteins, the pre-processing step must be completed for each of the proteins in the pair. The pre-processing step is generally run once for every protein in the species; the entire interactome (consisting of every possible pair) can then be predicted with PIPE. The run-time for the pre-processing step takes several orders of magnitude less time than the remainder of the PIPE algorithm and is therefore considered a flat start-up cost that does not significantly affect the overall runtime of PIPE.

### 3.2.2 PIPE Score Generation Algorithm

To predict whether a given pair of proteins, labelled ProtA and ProtB, interact, we first generate a score for the pair. If the score is greater than a threshold, we label it as an interaction. The score is generated by building a landscape of size length(ProtA)*length(ProtB) in size. This landscape is then reduced to a single value,

through an aggregation function such as the SW Score described in Section 2.3.3. This section will first analyze the generation of the landscape, as it is the most time-consuming step. This section will then briefly analyze the score generation algorithm.

A high-level overview of the algorithm used to generate the landscape is as follows: for every pair of windows in the two proteins, count how many interactions exist between the pre-computed similar proteins for each of the windows; then, fill in the landscape with this count. This was explained in more detail in Section 2.3.3 and Figure 10. The actual implementation of the algorithm to generate the landscape is shown in Figure 17.

```
    Input:
      Two proteins ProtA and ProtB
      Precomputed similarity files for each protein
      List of known interactions for every protein

    Output:
      landscape of size len(ProtA) * len(ProtB)

1   landscape = 2D array of size(len(ProtA) x len(ProtB))
2   all(landscape) <- 0

3   for each sliding window, WBx, in ProtB:

4     for each sliding window, WAy, in ProtA:

        // Look up similar proteins to WAy from similarity DB of ProtA
5       for each similar protein, SPz-WAy, in simprots(WAy):

          // Look up every protein that interacts with SPz-WAy
6         for each interaction partner, IP, in interactions(SPz-WAy):

            // Look up similar proteins to WBx from sim DB of ProtB
7           if IP is in simprots(WBx):
8             increment landscape at (y,x)
```

**Figure 17: Original PIPE algorithm for generating landscape.**

A Big O complexity analysis of the runtime behaviour of the landscape generation algorithm shown in Figure 17 can be used to analyze the expected impact of changes to the algorithm. The original Big O runtime is given by equation 3.1, which will be explained below. The line over each term is used to denote average

$$landscape\ generation = O(\overline{length(ProtA)} * \overline{length(ProtB)} * \overline{simProtsPerWindow} * \overline{interactionsPerProt}) \quad 3.1$$

Lines 3 and 4 of Figure 17 loop over every pair of sliding windows in Protein A and Protein B. Since the sliding windows are overlapping, the number of times each line is executed is proportional to the number of amino acids, or length of the two proteins. Since the lines are nested, all terms are multiplied together. This gives the term $\overline{length(ProtA)} * \overline{length(ProtB)}$. If protein A and B are of the same length on average, this term could be simplified to $\overline{length(Prot)}^2$. However the separate formulation is used throughout this chapter to account for inter-species PPI prediction, where Protein A and B may have different average lengths.

Lines 5-8 count the number of interactions between the similar proteins for each sliding window. Line 5 loops through all similar proteins to a sliding window from Protein A. The average number of similar proteins to a given window is denoted by $\overline{simProtsPerWindow}$. Line 6 loops through all the known interacting partners of a similar protein. The average number of interactions for a protein is denoted by $\overline{interactionsPerProt}$. Line 7 consists of an $O(1)$ check to see if a protein is a member of the set *simprots*($WB_x$). In a previous version of PIPE, line 7 instead consisted of a loop going through all proteins in *simprots*($WB_x$) to check if the interaction partner IP was in the set. However in [64], this loop was replaced with an $O(1)$ check by pre-processing the sparse similarity database of Protein B into a bit vector before the landscape was generated. Line 8 can also be completed in $O(1)$ time. Overall, this leads to the runtime given by equation 3.1.

The Big O runtime for the aggregation of the landscape into a single score and the pre-processing of Protein B to allow an $O(1)$ membership test are given below:

$$landscape\ summarization = O(\overline{length(ProtA)} * \overline{length(ProtB)}) \quad 3.2$$

$$preprocessing\ step\ = O(\overline{length(ProtB)} * \overline{simProtsPerWindow} \quad 3.3$$

As the above two equations are smaller than the landscape generation equation above, the summarization and pre-processing steps can be disregarded when estimating the overall Big O runtime of PIPE (as these terms are added, the smaller terms can be dropped in Big O formulation).

## 3.3 Modified PIPE Algorithm

Lines 5-8 of the PIPE algorithm explained in Figure 17 are used to compute the height of the proteinA-proteinB landscape at a given location (*y*,*x*). This algorithm can be reformulated as: *for window $WA_y$ from protein A and window $WB_x$ from protein B, count number of proteins appearing in <u>both</u> of the following two sets:*

$$IWSP_{Ay} = interacts\_with\left(simprots\left(WA_y\right)\right) \quad 3.4$$

$$SP_{\mathrm{Bx}} = simprots(\mathrm{WB_x}) \quad 3.5$$

The first set, denoted $IWSP_{Ay}$, contains the proteins that interact with the proteins that are similar to the window $WA_y$. The second set, denoted $SP_{\mathrm{Bx}}$, corresponds to the proteins similar to window $WB_x$. If a protein is in both sets $IWSP_{Ay}$ and $SP_{\mathrm{Bx}}$, then it represents a known interaction between a protein similar to window $WB_x$ and a protein similar to the window $WA_y$. In the original PIPE algorithm, this is done by looping through every member of $IWSP_{Ay}$ and performing an O(1) check whether the protein is in set $SP_{\mathrm{Bx}}$. As such, the Big O of finding the common subset is given by:

$$set\ intersection\ runtime = O(length(IWSP_{Ay})) \quad (3.6)$$
$$= O(\overline{simProtsPerWindow} * \overline{interactionsPerProt})$$

However, the probability that a specific protein in set $IWSP_{Ay}$ is also in set $SP_{Bx}$ is given by the following formula:

$$probability\ that\ any\ protein\ is\ in\ SP_{Bx} = \frac{length(SP_{Bx})}{total\_proteins} \approx \frac{\overline{simProtsPerWindow}}{total\_proteins} \quad (3.7)$$

where $total\_proteins$ is the total number of proteins in the species. Since $\overline{simProtsPerWindow}$ is generally much smaller than $total\_proteins$ (and by definition is always smaller or equal), when PIPE loops through the proteins of set $IWSP_{Ay}$, each protein will rarely be in set $SP_{Bx}$. This may seem unavoidable when only examining one pair of sets at a time. However, PIPE must repeat this process for every pair of windows in all pairs of proteins, resulting in many set intersection calculations. Therefore, we should consider an alternate data representation for PIPE.

Instead of finding the intersection of sets $IWSP_{Ay}$ and $SP_{Bx}$ for each pair of sliding windows in the input proteins, we propose an alternate data representation in which each protein (or each potential member of the set intersection between $IWSP_{Ay}$ and $SP_{Bx}$) is a key or index, and the sliding windows it occurs at within each input protein is the value. Instead of keeping the set of proteins $SP_{Bx}$ for each window location *x* in protein B, we instead store $SP_{ProtID}$, which is the set of all locations, *x,* for which the set $SP_{Bx}$ contains *ProtID*. Similarly, we also transform $IWSP_{Ay}$ to $IWSP_{ProtID}$, which is the set of all locations *y* for which the set $IWSP_{Ay}$ contains *ProtID*. Essentially, rather than list all proteins similar to a given window, list all windows similar to a given protein. This way, we can build the common subset (and the interaction landscape) for every pair of windows at once by

examining each *ProtID* and incrementing the landscapes for every pair of windows between the sets $SP_{ProtID}$ and $IWSP_{ProtID}$. We would therefore never perform a membership check that returns false, and we expect to be able to reduce the time to build the landscape by the following factor:

$$\frac{O(t_{LandscapeGenerationOld})}{O(t_{LandscapeGenerationNew})} = \frac{total\_proteins}{simProtsPerWindow} \quad 3.8$$

### 3.3.1 Modified Input Data Representation

The new method of data representation requires two different types of database files for each protein. This first type of database file, referred to as a Type 1 DB file, is meant to represent the list of sets of proteins similar to a given window, denoted as $SP_{Bx}$ in equation 3.5 above (we refer to "list of sets" instead of "set" since we are now considering the set for every window within a protein instead of just a single window). This database file is organized to by *ProtID* to represent the $SP_{ProtID}$ transformation. This first database file is essentially the original database file (shown in Figure 16) flipped: instead of storing the similar proteins for each window (of the database protein), the new database stores the similar windows (of the database protein) for each known protein. The new database format is shown in Figure 18. Note that we know use simW(P#) instead of simP(W#) to reflect this difference.

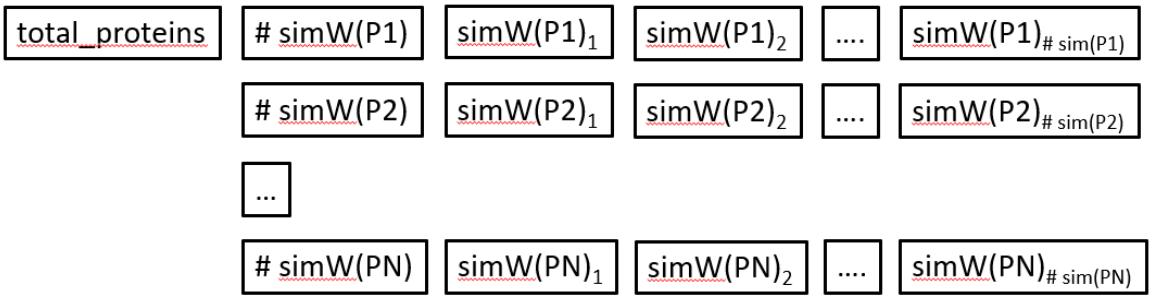



**Figure 18: Modified database format example for a single protein, Type 1 DB file. For each protein, store similar windows (simW) locations (e.g. amino acid offset) for each protein**

The modified pre-processing algorithm is shown in Figure 19. It is very similar to the original pre-processing algorithm in Figure 15. The only real difference is that lines 2 and 4 were swapped in the newer algorithm in order to keep track of similar windows for each protein instead of similar proteins for each window.

```
     Input
      list of proteins and sequences

     Output
      Type 1 DB file for each protein

1    for each protein, p1 in list of proteins:

2      for each protein, p2 in list of proteins:

3        let similarWindows = []
4        for each sliding window, W1x, in p1:

5          let p2hasAnySimilarx = False
6          for each sliding window, W2y, in p2:

7            if similarity(W1x,W2y) > threshold:
8               set p2hasAnySimilarx = True

9          if p2hasAnySimilarx == True:
10            add W1x to similarWindows

11       write a new line of p1 Type 1 DB file with
     len(similarWindows), similarWindows
         // this contains list of windows in p1 that are similar to p2
```

**Figure 19: Modified pre-processing algorithm for PIPE to create Type 1 DB files.**

The second type of database file, referred to as Type 2 DB file, is meant to represent the list of sets denoted as $IWSP_{Ay}$ in equation 3.4 above. This database file is organized to by *ProtID* to represent the $IWSP_{ProtID}$ transformation. That is, a Type 2 DB file for protein A lists, for each protein (labelled *ProtID*), the list of window locations in protein A that are similar to a protein that interacts with the protein *ProtID*. The Type 2 DB file structure is shown in Figure 20. Note that this database file can sometimes contain the same window multiple times per protein (for example if a window in the database protein is similar to multiple proteins that interact with same protein). The Type 2 DB files are generated from the Type 1 DB file and a list of known interactions by following the algorithm given in Figure 21.

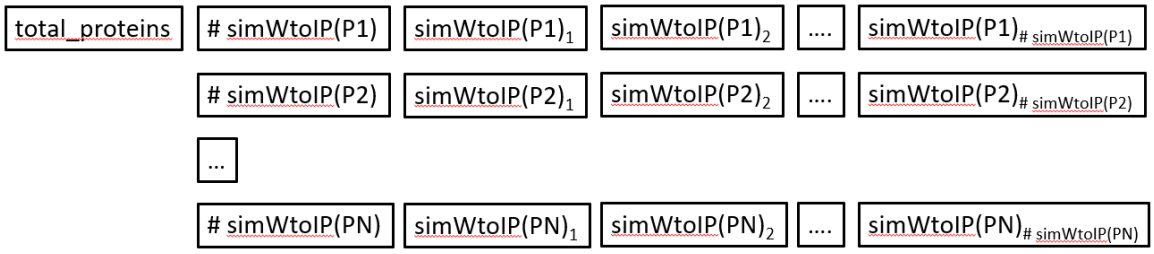



**Figure 20: Modified database format example for a single protein, Type 2 DB file. For each protein, list all windows (of protein A) that are similar to a protein that is an interaction partner (simWtoIP) with that protein.**

```
     Input
       List of Type 1 DB files
       List of known protein-protein interactions

     Output
       New Type 2 DB file for each protein

1    helper = array of size total_proteins
     // each array index contains a linked list to store proteins

2    for each protein, p1 in list of proteins:
3      db1 = Type 1 DB file for p1

4      for each protein, p2 in list of proteins:
5        simwindows = read similar windows from db1 for protein p2
         // p2 line of contains list of windows similar to p2

6        for each known interaction partner, intpartner, of protein p2:
7          add simwindows to helper[intpartner]


8      for each protein, p2 in list of protein:
9        write to p1 Type 2 DB file with length(helper[p2]), helper[p2]
         // this contains list of windows in p1 that are similar
         // to proteins that interact with p2
```

**Figure 21: Modified pre-processing algorithm for PIPE to create Type 2 DB files.**

As mentioned previously, the pre-processing step of PIPE takes significantly less time than the prediction step, and can be considered a flat, start-up cost. The change to the pre-processing step outlined in this section did not change the runtime of the pre-processing step in any notable way, as the pre-processing step time complexity is dominated by the amino acid comparison step (lines 4-8 in Figure 19). Since the number of amino acid comparisons is necessarily the same in both algorithms, the expected runtime does not change.

### 3.3.2 Modified PIPE Score Generation Algorithm

Using the Type 1 and 2 DB files introduced above, the new PIPE algorithm is given in Figure 22 below. The new algorithm will produce the same results as the original algorithm explained in Figure 17 above, but with different computational cost.

```
     Input:
       Two proteins ProtA and ProtB
       Precomputed Type 1 and Type 2 DB files

     Output:
       landscape of size len(ProtA) * len(ProtB)

1    landscape = 2d array of size(len(ProtA) x len(ProtB))
2    all(landscape) <- 0

3    db1 = read Type 1 DB file for ProtA
4    db2 = read Type 2 DB file for ProtB

5    for each protein, protID, in list of proteins:
6       simwindowsA = read protID entry from db1
7       simwindowsB = read protID entry from db2

8       for every window, simWA in simwindowsA:
9         for every window, simWB in simwindowsB:
10              increment landscape at (simWA, simWB)
```

**Figure 22: Modified PIPE landscape generation algorithm.**

The Big O computational cost of the new algorithm is given by the following equation:

$$new\ landscape\ generation = O\left(numProteins * \overline{len(d1PerProt)} * \overline{len(db2PerProt)}\right) \quad 3.9$$

Where $\overline{len(db1PerProt)}$ is the average number of similar windows for a protein in a Type 1 DB file and $\overline{len(db2PerProt)}$ is the average number of similar windows for a protein in the type 2 database file. The Big O of the new landscape generation algorithm can be further simplified by substituting the following equations, which can be found by analyzing the algorithms given in Figure 19 and Figure 21.

$$\overline{len(db1PerProt)} = \frac{\overline{length(ProtA)} * \overline{simProtsPerWindow}}{total\_proteins} \quad 3.10$$

$$\overline{len(db2PerProt)} = \frac{\overline{length(ProtB)} * \overline{simProtsPerWindow}}{total\_proteins} * \overline{interactionsPerProt} \quad 3.11$$

Substituting these into equation 3.12, we find the new Big O to be:

$$new\ landscape\ generation = O\left(\frac{\overline{length\ (ProtA)} * \overline{length(ProtB)} * \overline{simProtsPerWindow}^2 * \overline{interactionsPer\ Prot}}{total\_proteins}\right) \quad 3.12$$

Comparing this equation to the original landscape generation equation (equation 3.1 in section 3.2.2) we find:

$$new\ landscape\ generation = O\left(original\ landscape\ generation * \frac{\overline{simProts\ PerWindow}}{total\_proteins}\right) \quad 3.13$$

The above equation confirms our assertion in Section 3.3 about how a new algorithm could avoid checking proteins that will not be in both sets. Since $\overline{simProtsPerWindow}$ must be smaller than $total\_proteins$, we expected the new landscape generation to be faster. In practice, $total\_proteins$ can be several orders of magnitude larger than $\overline{simProtsPerWindow}$, resulting in significant acceleration of the new PIPE algorithm.

## 3.4 Benchmarking Methodology and Data

The original PIPE algorithm, written in the C programming language using MPI and OpenMP for parallelization, was modified with the changes previously explained. Since PPI prediction is "embarrassingly parallel", the total algorithm runtime is most affected by the time to calculate a score for a given protein pair. The C programming language is therefore expected to produce the fastest implementation for either algorithm. After showing that the new and old algorithms produced the same score for a given protein pair, the two algorithms were then timed and compared on several different species to examine the change in runtime. The species tested on were *H. sapiens, A. thaliana, S.*

*cerevisiae*, three important model species previously examined with PIPE. The prediction of *H. sapiens* interactions has historically taken the longest of all model species tested with PIPE, while *S. cerevisiae* has taken the shortest and *A. thaliana* has been in between. The new and old PIPE algorithms were timed on the prediction of a subset of protein pairs for each species. Every known interaction was predicted (labelled positive in table), and a subset of random protein pairs were tested (labelled negative in table). The sequences for each species were taken from UniProt [68], and known interactions were taken from BioGRID [20]. The testing data, as well as different parameters referenced previously are shown in Table 1 below, along with the number of pairs that would be tested for a full interactome prediction.

**Table 1: Intra-species PPI prediction test data for comparing new and old PIPE implementation algorithms.**

| | *H. sapiens* | *A. thaliana* | *S. cerevisiae* |
|---|---|---|---|
| $\overline{simProtsPerWindow}$ | 35.1 | 21.7 | 8.2 |
| $\overline{length(Prot)}$ | 558 | 442 | 450 |
| $\overline{interactionsPerProt}$ | 6.45 | 3.33 | 8.15 |
| $total\_proteins$ | 20,236 | 17,226 | 6,721 |
| Total known interactions | 66,084 | 29,035 | 27,905 |
| Number positives tested | 66,084 | 29,035 | 27,905 |
| Number negatives tested | 66,084 | 3,000,000 | 3,000,000 |
| Number of pairs in all-to-all | 204,757,966 | 148,376,151 | 22,589,281 |

Afterwards, the modified algorithm was used on two different inter-species predictions of interest to Agriculture and Agri-Food Canada. The two inter-species predictions were between *H. glycines and G. max* (SCN – soybean), and between *H.*

*sapiens and G. max* (human – soybean). The species and interactions used to make these predictions are given in Table 2 below. The reasons for using the different training species will be further explained in Chapter 5. For the purposes of this chapter, it is sufficient to consider inter-species prediction as simply pooling together interactions and sequences from multiple species to create a single 'pooled species' on which to run PIPE. Since these predictions were of interest to Agriculture and Agri-Food Canada, all possible protein pairs were tested. The new implementation was used to complete an all-to-all prediction for each inter-species case. The old implementation was run in full on the SCN – soybean test, but only on a portion of interactions from the human – soybean test.

**Table 2: Inter-species predictions completed with PIPE at the behest of Agriculture Canada.**

| | *H. sapiens* - *G. max* | *H. glycines* - *G. max* |
|---|---|---|
| $\overline{simProtsPerWindow}$ | *H. sapiens*: 71.9<br>*G. max*: 89.1 | *H. glycines*: 104.1<br>*G. max*: 77.8 |
| $\overline{length(Prot)}$ | *H. sapiens*: 558<br>*G. max*: 336 | *H. glycines*: 406<br>*G. max*: 336 |
| $\overline{interactionsPerProt}$ | 1.66 | 0.287 |
| $total\_proteins$ | *H. sapiens*: 20,236<br>*G. max*: 75,781<br>A. thaliana: 17,226<br>Total: 113,243 | *G. max*: 75,781<br>*H. glycines:* 21,868<br>*C. elegans:* 26,725<br>*A. thaliana*: 17,226<br>Total: 141,600 |
| Number of known interactions | *H. sapiens – H. sapiens*: 66,084<br><br>*A. thaliana – A. thaliana:* 29,035<br><br>Total: 95,119 | *C. elegans – C. elegans*: 5,476<br><br>*A. thaliana – A. thaliana:* 29,035<br><br>Total: 34,511 |
| Total_pairs to predict | 1,594,212,316 | 1,657,178,908 |
| Total_pairs tested w/ new | 1,594,212,316 | 1,657,178,908 |
| Total_pairs tested w/ old | 143,992,932 | 1,657,178,908 |

All testing runs were completed on a local Carleton University computing cluster, comprising 18 compute nodes, each with a 100 GB SSD, 32 GB of RAM, and an Intel Core i7-3770 8-core processor at 3.40 GHz. For the larger all-to-all predictions, all 18 nodes were utilized to their full capacity, without any other programs being run (with the exception of programs run by the operating system). For the inter-species prediction, only 2 threads could be executed in parallel due to the lack of available memory. For the intra-species predictions, all 8 threads could be used. The run times of the different algorithms were found by instrumenting the code to report the run time of each protein pair tested; the average of these times is reported. While there would be some background noise due to operating system programs and measurement noise, they are expected to balance out over the various tests and protein pairs tested. The wall time of the entire testing set was compared to the average time per protein pair times the total protein pairs tested divided by the number of threads, and the results were very similar. This demonstrates that average wall time per protein pair is useful metric for comparing algorithms.

In addition to the time taken to run PIPE on each protein pair, the times to generate the landscapes were also measured. These times also reflect the maximum potential speed up with PIPE. If only the SW score is calculated for each protein pair, then only the landscape generation section of the algorithm needs to be run. This is because the SW score calculates the overall score by taking the mean of each landscape value (modified by similarity values). The mean can be tracked by storing the total sum and number of window pairs, without storing the complete landscape. The other scoring methods use a filter, and therefore need the entire landscape. By not storing a complete landscape, PIPE can be run slightly faster while still calculating the SW score. This would also result in PIPE using

significantly less memory, potentially allowed more threads to be launched simultaneously. Since the SW score is the main score using, the landscape generation times give the maximum possible speedup with the new algorithm.

## 3.5 Results and Discussion

Table 3 shows the database size, database processing time, and PPI prediction timing results for the three intra-species PPI prediction tests. It can be seen that the database size increases with the new algorithm, but never increases passes the 100 GB limit on the local cluster used. The modified algorithm represents a space-time trade-off in the algorithm: more data is stored for each protein, but the computation of the landscape is performed more quickly. Since the database is stored as individual files for each protein on a SSD, the only limitation is that the database must fit on the SSD. Since this limit is not reached, the only significant effect of the space-time trade-off is an increase in computational speed. The database pre-processing also takes a very similar length of time with both algorithms. For each species, the new algorithm is significantly faster with speedups ranging from 8.3 times faster for *S. cerevisiae* to 53.2 times faster for *H. sapiens*. Additionally, the modified algorithm has the largest improvements on the species that take the longest time per protein pair. The all-to-all prediction times are also extrapolated to illustrate the full impact of the change when running PIPE.

**Table 3: Intra-species time comparison tests for new and old PIPE algorithm implementation. All-to-all and database processing time are measured while using the full resources of the local cluster.**

| | *H. sapiens* | | *A. thaliana* | | *S. cerevisiae* | |
|---|---|---|---|---|---|---|
| | New | Old | New | Old | New | Old |
| Database size (GB) | 18 | 1.6 | 5.6 | 0.7 | 1.9 | 0.2 |
| Database processing time (s; using 18 nodes, 8 threads/node) | 3191 | 3194 | 1358 | 1325 | 263 | 255 |
| Time/pair positive (s) | 0.0155 | 0.7700 | 0.0061 | 0.1103 | 0.0113 | 0.1280 |
| Time/pair negative (s) | 0.0084 | 0.4447 | 0.0049 | 0.0615 | 0.0054 | 0.0448 |
| Weighted speed (s; with 500:1 CI) | 0.0084 | 0.4453 | 0.0049 | 0.0616 | 0.0054 | 0.0449 |
| All-to-all prediction time estimate (hours; 500:1 CI; using 18 nodes, 8 threads/node) | 3.3 | 175.9 | 1.4 | 17.6 | 0.2 | 2.0 |
| Landscape generation (s; weighted speed at CI 500:1) | 0.0056 | 0.4405 | 0.0022 | 0.0586 | 0.0029 | 0.0427 |
| **Total Speedup (~x faster)** | **53.2x** | | **12.5x** | | **8.4x** | |
| **Landscape generation Speedup (~x faster)** | **79.2x** | | **26.4x** | | **14.8x** | |

The new algorithm is comparable to another state-of-the-art algorithm, SPRINT [38]. The SPRINT method is reportedly able to predict the entire human interactome in approximately 11 hours on a specific 12-core machine. The new version of PIPE is able to perform this task in approximately 40 hours using 12 threads (on a different but comparable computer), which is comparable to SPRINT. Additionally, if the PIPE score landscape is not explicitly computed and the PIPE score is instead calculated directly, then PIPE's runtime is reduced to 26 hours. This shows that the time required for PIPE is now at least within an order of magnitude of the state of the art. For comparison, the original version of

PIPE would take approximately 2110 hours (88 days) to perform the same analysis on a 12 core machine.

In the case of inter-species PPI prediction, similar speedups are seen as shown in Table 4. With the new algorithm, a prediction between all human proteins and all soybean proteins was able to be completed within ~3 days. Previously, this would have taken ~72 days using the local cluster. The prediction between SCN and soybean was also completed using the revised algorithm in ~2 days instead of the ~42 previously required.

**Table 4: Inter-species time comparison tests for new and old PIPE algorithm implementation. All-to-all and database processing times reflect usage of all cluster resources. All-to-all predictions times were estimated using time/pair for consistency, but were similar to the actual total times.**

| | *H. sapiens* - *G. max* | | *H. glycines* - *G. max* | |
|---|---|---|---|---|
| | New | Old | New | Old |
| Database size (GB) | 69 | 12 | 66 | N/A* |
| Database processing time (on 18 total threads) (h) | 56.33 | 56.30 | 44.01 | 44.00 |
| Time/pair (s) | 0.0057 | 0.1410 | 0.0037 | 0.0785 |
| All-to-all prediction time (using 2 threads on 18 nodes) (hours) | 69.6 (2.9 days) | 1733.9 (72.2 days) | 47.6 (2.0 days) | 1003.8 (41.8 days) |
| **Speedup (~x faster)** | **24.9x** | | **21.1x** | |

* unfortunately, the original database files for *H. glycines* - *G. max* were deleted due to lack of space on local cluster so their size cannot be reported.

While substantial, the improvement in speed did not match the projected improvement from Big O analysis. This was expected however, as Big O does not reflect all constants that affect runtime.

## 3.6 Conclusions

This chapter described the acceleration of the PIPE algorithm by changing the data structures created during the pre-processing step. The changes leverage the fact that many set intersection calculations must be performed to generate the PIPE score. While the original PIPE algorithm used an optimal algorithm for set intersection calculation between two sets, the modified algorithm transformed the list of sets data structure so that every set intersection calculation could be found directly, without needing to examine every set element like in the original algorithm. Experimental tests were used to verify that the modified algorithm was significantly faster for many common PIPE use cases, and is now comparable to the time of the leading state-of-the-art method, SPRINT. The new algorithm did not significantly change the pre-processing time or memory usage required for PIPE. The modified version will be contributed back to the Carleton Bioinformatics Group and used for future PIPE runs.

The new PIPE algorithm raises a number of issues. One issue is the increase in database file size. The database file size did not affect runtime on the test cluster, as the largest set of database files was 69 GB, well below the SSD capacity of 100 GB. However, if PIPE were run on different clusters or for different species, the increase in size may be prohibitive. Another issue is that the pre-processing step must be re-run if the known interactions list changes. This makes it more difficult to use different list of known interactions, for example testing PPI prediction using interactions found using different experimental methods. For the same reason, the new algorithm cannot be used in LOOCV testing, since LOOCV requires rerunning PIPE with slightly different interaction lists each time. This makes it more difficult to optimize the score decision threshold for intra-species

PPI predictions. However, it would still be possible to use the original PIPE algorithm for LOOCV and then the new algorithm for all-to-all predictions. The database-pre-processing would have to be re-run, but a significant amount of time would be saved for larger all-to-all predictions.

# 4 Cross-species PPI Prediction

## 4.1 Introduction

PIPE has previously shown success in predicting intra-species PPI in relatively well-studied species with large numbers of experimentally verified PPI. Cross-species PPI prediction, where PPI from one species can be used to predict PPI in another species, is another important use case. Cross-species PPI prediction allows experimental data taken from well-studied species to be used to learn more about under-studied species, which often have very little data available to them, but are still important topics for research. For example, to predict PPI in SCN and soybean, which are both under-studied species with few known PPI, we will leverage cross-species prediction using *C. elegans* and *A. thaliana* model organisms respectively.

A preliminary demonstration of PIPE's ability to complete cross-species PPI prediction was presented in [11]. This work showed that the PPI in Baker's yeast (*S. cerevisiae*) could be used to predict human (*H. sapiens*) PPI and vice-versa. However, they only showed that the predictions were better than random chance, and did not explore or develop best practice for making cross-species predictions. In this chapter, best practices for cross-species PPI prediction will be examined, including from which species to take training data, for which species valid predictions can then be made, whether using combinations of training species could be advantageous, and whether the PIPE score needs to be modified for cross-species prediction.

## 4.2 Methods

### 4.2.1 Training Data

In order to test the cross-species prediction performance of PIPE, data from multiple species with known interactions was needed. Therefore, the PPI data from all available model species was downloaded from BioGRID [20]. The data were filtered to only include physical PPI, and not genetic interactions or functional associations. The types of interactions in BioGRID were listed as physical association, direct interaction, colocalization, and association. The data were then filtered to only include intra-species interactions (in order to simulate using one species' intra-species interactions to predict another species' intra-species interactions). Duplicate interactions were removed from the data. The data were then filtered to exclude species that had ten or fewer PPI. There were seventeen species that met this criterion. For each of those species, the amino acid sequences of every protein was downloaded from the UniProt database [68]. The Swiss-Prot database is the subset of proteins that have been manually annotated and reviewed by UniProt. If a known interaction from BioGRID did not correspond to a Swiss-Prot protein, then the sequence instead was taken from the larger TrEMBL UniProt database, which comprises automatically annotated and reviewed proteins. The results of this data processing is shown in Table 5 below.

**Table 5: Number of intra-species PPI for the different species used, sorted by number of interactions.**

| Scientific name | Common name | # sequences | # interactions |
|---|---|---|---|
| *Homo sapiens* | Human | 20,236 | 66,084 |
| *Arabidopsis thaliana* | Mouseear cress | 17,226 | 29,035 |
| *Saccharomyces cerevisiae* | Brewer's yeast/ Baker's yeast | 6,721 | 27,905 |
| *Drosophila melanogaster* | Common fruit fly | 8,529 | 25,013 |
| *Caenorhabditis elegans* | Nematode (roundworm) | 5,891 | 5,476 |
| *Schizosaccharomyces pombe* | Fission yeast | 5,141 | 3,549 |
| *Mus musculus* | House mouse | 17,096 | 3,402 |
| *Plasmodium falciparum* | Malaria parasite | 1,270 | 2,250 |
| *Rattus norvegicus* | Common rat | 8,129 | 531 |
| *Xenopus laevis* | African clawed frog | 169 | 117 |
| *Solanum lycopersicum* | Tomato | 474 | 98 |
| *Danio rerio* | Zebrafish | 3,080 | 94 |
| *Oryza sativa* | Asian rice | 3,832 | 29 |
| *Gallus gallus* | Red junglefowl (chicken) | 2,305 | 28 |
| *Bos taurus* | Bovine (cow) | 6,026 | 26 |
| *Glycine max* | Soybean | 429 | 24 |
| *Candida albicans* | Pathogenic yeast | 1,014 | 19 |

### 4.2.2 Normalization Factor Change

Normalization of the interaction landscape (as explained in Section 2.3.3) by sequence window uniqueness has shown great success for same-species PPI prediction in past [64], and is the main way the landscape is summarized into a single SW-score for a protein pair by PIPE (after this normalization, the score for the protein pair is simply the average value of the landscape). However, this basic form of normalization may not be suitable for cross-species PPI prediction. Consider the example of using known mouse-mouse and human-human intra-species training PPI (which are numerous as they are very well studied) to predict intra-species PPI in the test species dog (of which very few are known). Using the original normalization approach, for a given pair of windows, WA and

WB, the count of known interactions would be normalized by the total number of possible interactions as in equation 4.1:

$$normfactor = (f_{human}^{WA} + f_{mouse}^{WA} + f_{dog}^{WA}) * (f_{human}^{WB} + f_{mouse}^{WB} + f_{dog}^{WB}) \quad 4.1$$

Where, $f_{human}^{WA}$ is the frequency of window A in human and so on. This equation implies that there are possible interactions between proteins from human and mouse, and normalizes as such. While human and mouse proteins may interact in-vitro, no such inter-species interactions are used in the training data. This makes the normalization factor unfairly penalize amino acid windows when there are similar windows in multiple species. This effect will become more pronounced as the number of similar proteins and the number of species increases. Additionally, it also normalizes by how frequently the window shows up not only in the training species, but also in the test species dog proteome. Since very few interactions are known for dog (and none used for training), this normalization factor does not reflect the true number of possible interactions among similar proteins in the training set, and should not be accounted for.

The new normalization factor (seen in equation 4.2) takes this into account, and only normalizes between species when there is training available data for that species. This is illustrated in Figure 23 below.

$$normfactor = f_{human}^{WA} * f_{human}^{WB} + f_{mouse}^{WA} * f_{mouse}^{WB} \quad 4.2$$

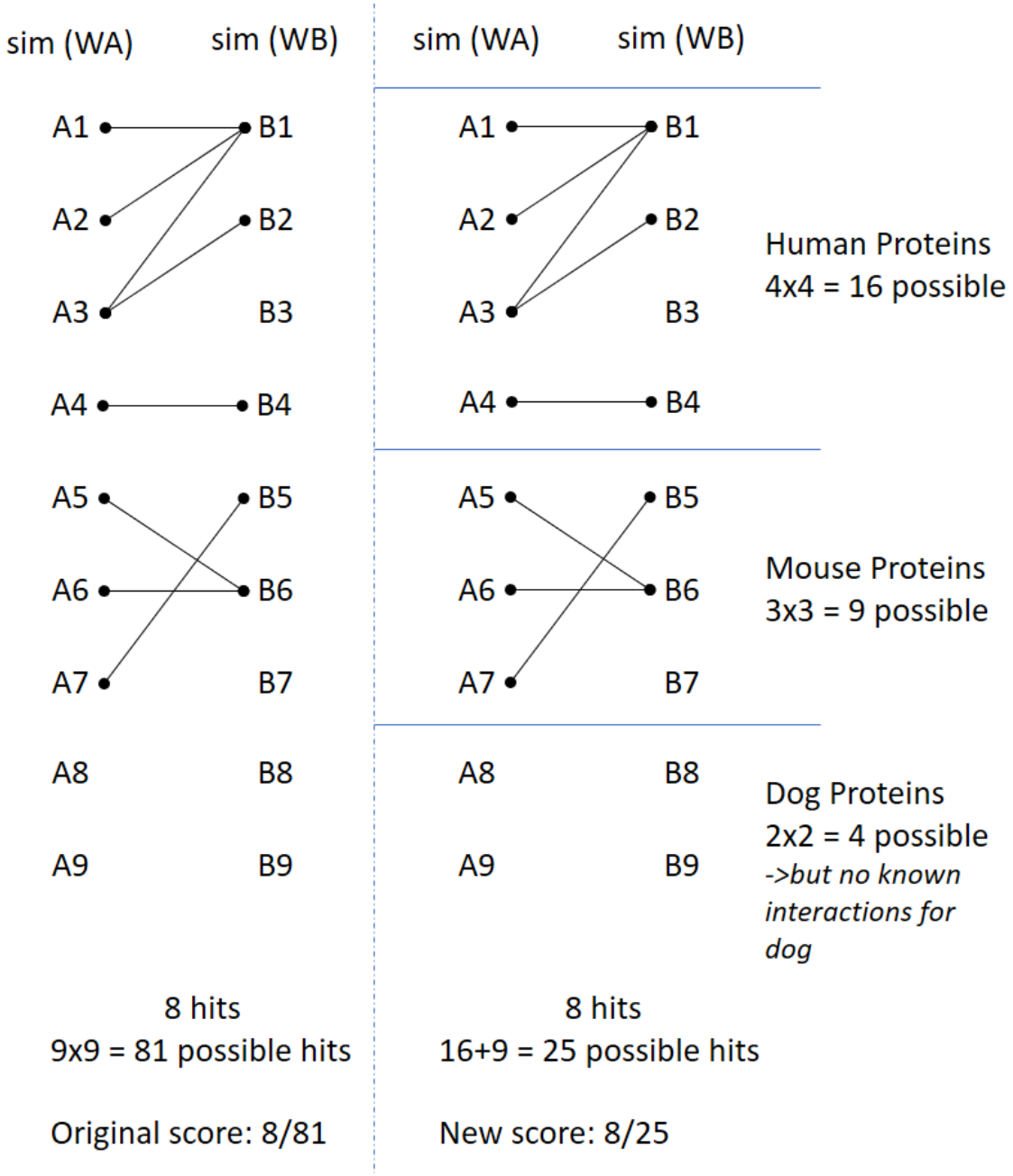



**Figure 23: SW score normalization change for cross-species prediction. In this example, dog-dog PPI are predicted from human-human and mouse-mouse training data.**

A cross-species version of PIPE was created to account for this. The new version of PIPE keeps track of the species of origin for each protein, and normalizes the score by only the relevant type of interaction available in the training data. For example, if human-mouse PPI were available, the normalization factor in Figure 23 would increase to reflect

the 7x7=49 possible interactions between human-human, mouse-mouse, and human-mouse proteins.

Figure 24 illustrates how the normalization factor would change as the number of species and number of similar proteins changes. In this figure, the number of proteins on the x-axis represents how many similar proteins are seen in each species for a single window. As the number of species increases, the effect of the normalization change becomes more pronounced. Furthermore, since each protein window has a different number of similar proteins, the change in normalization factor affects each PIPE prediction differently; it is not strictly a uniform scaling across all predicted interactions.

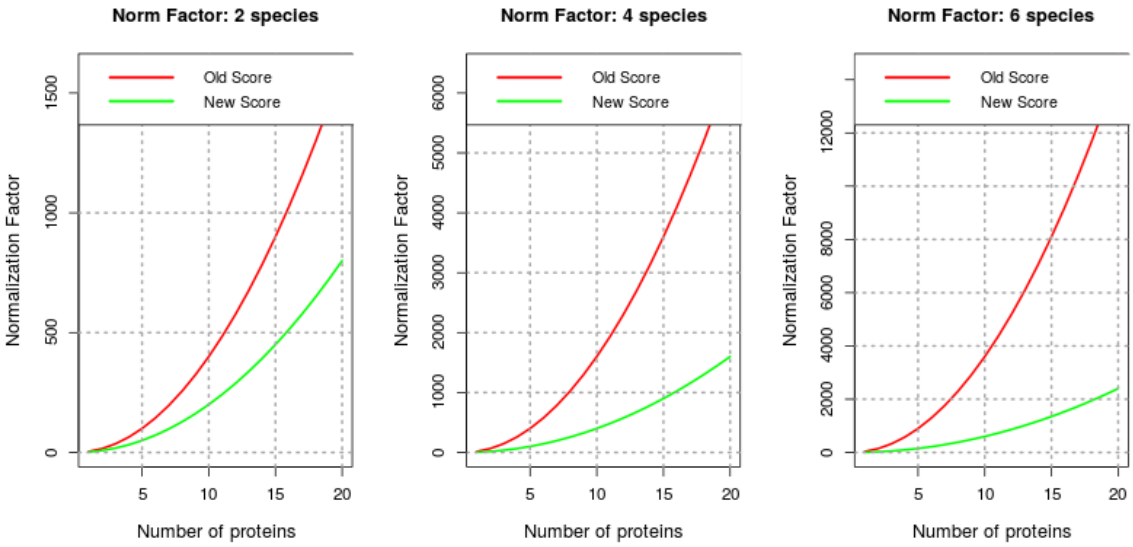



**Figure 24: Normalization factor as a function of the number of proteins for 3 different numbers of species. The number of proteins number represents how many similar proteins are seen in each species for each of the two windows being examined.**

The effect of this normalization change was tested using the species data given in Section 4.2.1. The eight species with more than 2000 known interactions were used for this experiment. Two different tests were completed. The first test used each of the eight species to predict the seven remaining species using cross-species predictions. This test examined

if the score normalization had any effect for cross-species predictions involving a single training species. ROC and PRC were generated, and the area under the PRC (AU-PRC) and precision at 25% TPR (Pr@25%TPR) were used as scoring metrics to compare the classification performance of the original score and the modified score. The second test made predictions in each species using all seven remaining species at once. This tested whether the normalization change becomes more pronounced as the number of training species increases. No sub-sampling of the interactions was done at any step; all available interactions were used for the cross-species predictions. In order to test if the normalization factor change improved the performance on the two metrics given, a Student's paired t-test was performed using the R programming language. The paired t-test function produces an estimate of the true difference in the means of the performance metrics for each normalization factor, as well as a p-value indicating how likely this difference is under the null hypothesis that the means are the same.

### 4.2.3 Evolutionary Distance Relation to Classification Performance

A main hypothesis tested in this chapter was that a training species that is more closely related evolutionarily to a test species would perform better for cross-species PPI prediction, as protein sequence, structure, and function are more likely to be conserved between closer species. To test this, the evolutionary divergence times were taken from TimeTree [69] (available at www.timetree.org). The evolutionary divergence timeline estimates are shown in Figure 25 below for all species. Note that multiple species may be equidistant from a given species. For example, *M. musculus* and *R. norvegicus* are equidistant from *H. sapiens*.

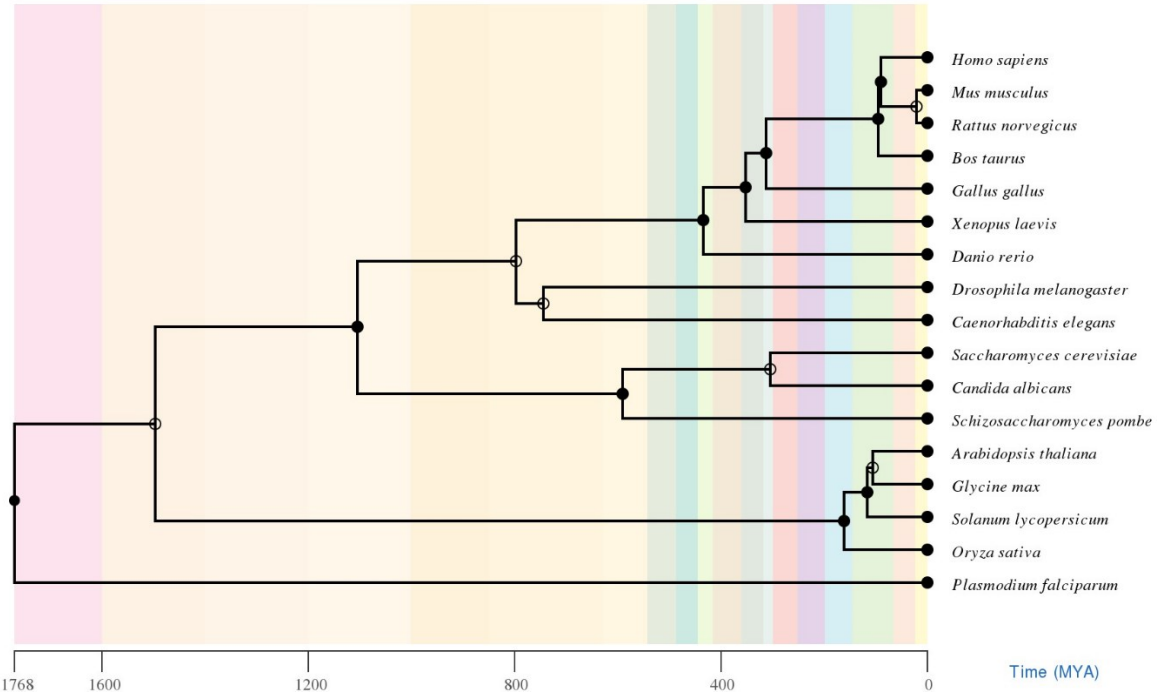



**Figure 25: Evolutionary divergence timeline estimates for species used (generated with** [69]**). The units are million years ago (MYA). The colours represent geological era and can be disregarded.**

To test if evolutionary distance is related to accuracy, we ran tests to predict every species' PPI from every other species' PPI. The hypothesis being that the performance ranks would be inversely correlated with evolutionary distance ranks. For example, when predicting *M. musculus* (mouse) PPI, the highest performance would be expected when training with *M. musculus* PPI, followed by training with *R. norvegicus* (rat) PPI, followed by *H. sapiens* (human) PPI and so forth.

However, there are different amounts of known interactions for each species; it may be that using a better-studied species with more interactions for training would perform better even though it is further away evolutionarily. To control for this, we ran a test controlling for size of the known interactome. Limiting ourselves to the eight species with more than 2000 interactions, we predicted the entire interactome of each of the eight species from 2000 randomly sampled interactions from each of the other eight species. We

tested the performance of the cross-species prediction on all known positive interactions of the test species as well as 100,000 random pairs from the test species to be used as a negative set. This allowed us to create a ROC curve for each test species from each training species. To prevent the random sampling of interactions from affecting results significantly, we repeated the entire process 20 times drawing different training subsets each time. The 20 different ROC curves for each test-training species pair were then averaged to summarize the performance of any training species on any testing species.

We used Kendall's Tau-b rank correlation test and the Spearman rank correlation test to examine if the relative evolutionary distance (i.e. ranking training species in order of most recent common ancestor with testing species) was correlated to the experimental performance ranks for each training species. For example, in predicting *H. sapiens* PPI, the evolutionary distance ranks of the other species are given in Table 6 below, alongside hypothetical ranks on a performance metric that are perfectly correlated to the expected (though without tied performance ranks). Note that species that are equally divergent from the test species are given the same evolutionary distance rank. The correlation coefficients would be calculated on the second and third columns in Table 6. Since there are multiple species that are equally distant evolutionarily, a perfect correlation is unlikely.

The performance metrics tested were area under the precision-recall curve (AU-PRC) and precision at 25% TPR (Pr@25%TPR). When computing precision, a class imbalance of 10:1 (ten negatives for every positive) was used. Although changes in class imbalance could affect relative order for the AU-PRC metric, they would not affect the relative ordering of the Pr@25%TPR metric.

**Table 6: Hypothetical performance rank resulting from using 8 different training species to predict *H. sapiens* PPI. The evolutionary distance ranks give the relative order amongst the training species for the degree of closeness between the training species and the test species *H. sapiens*. The performance ranks column gives the ordering from best training species to worst training species for the performance metric chosen.**

| Training species | Performance ranks | Evolutionary Distance Ranks |
|---|---|---|
| *H. sapiens* | 1 | 1 |
| *M. musculus* | 2 | 2 |
| *C. elegans* | 3 | 3 |
| *D. melanogaster* | 4 | 3 |
| *S. pombe* | 5 | 4 |
| *S. cerevisiae* | 6 | 4 |
| *A. thaliana* | 7 | 5 |
| *P. falciparum* | 8 | 6 |

To test the statistical likelihood of a correlation between evolutionary distance and prediction accuracy, we used the null hypothesis that there was no relation between evolutionary distance and accuracy. This is equivalent to predicting the Spearman and Tau-b correlation coefficients to be zero for data like seen in Table 6. The probability of observing the experimental results under the null hypothesis (i.e. the p-value) was found in two different ways. The first was a statistical method found through approximations of the spearman and Tau distributions. This method was implemented in the R programming language based on [70]. The second method of generating a p-value was through permutation tests. In this method, the Spearman and Kendall's correlation coefficient are calculated for each training species on a test species. The evolutionary distance ranks (as seen in Table 6) are then shuffled randomly many times, and the correlation coefficients are calculated each time. The p-value is then the percentage of shuffled ranks which produce a correlation coefficient as extreme as the experimentally observed value. These

two p-value methods were performed on each performance metric for every pair of training and testing species.

The previous methodology produces a p-value for each test species. It would be possible for some test species to have significant p-values while other test species will not have significant p-values. In order to summarize results over all the different test species, the experimental data should also be combined in a single test. This was done with a modified form of the Kendal Tau-b rank correlation. Kendall's Tau-b formula is given in the equation 4.3, from [70].

$$\tau_B = \frac{n_c - n_d}{\sqrt{(n_0 - n_1)(n_0 - n_2)}} \tag{4.3}$$

In this equation, $n_c$ and $n_d$ represent the number of concordant and discordant pairs of ranks. This is found by looping through every pair of paired ranks (e.g. every pair of rows in Table 6) and counting the number of times when both the expected and actual ranks are homogenously larger or smaller in one row or the other (concordant ranks) and the number of times this does not occur (discordant ranks). The divisor of the equation is a normalization correction to account for the number of non-duplicate rank pairs. $n_0$ is the total number of rows in the table, while $n_1$ and $n_2$ are the number of duplicate ranks for each column. This correlation coefficient was adapted by counting the number of concordant and discordant ranks for each test species individually before summing the results and calculating a single modified Tau-b coefficient that encompasses every pair of training and testing species. Since the modifications may invalidate the assumptions used in regular p-value estimation for the Tau-b test, the p-value was found through permutation tests only. Once again, the permutation test involved shuffling the evolutionary distance ranks for each species repeatedly, calculating the modified Tau-b coefficient for each

shuffle, and then finding the percentage of times when the shuffled correlation coefficient was as large as the experimental correlation coefficient.

## 4.3 Results

In this section, results are presented for two types of cross-species experiments. First, the new PIPE SW-scoring normalization change is examined for both a single training species and for multiple training species. Second, the classification performance as a function of the evolutionary distance between the training and testing species is examined.

### 4.3.1 Normalization Factor Accuracy Difference

#### 4.3.1.1 Single Cross-Species

Figure 26 shows the ROC curves for predicting *H. sapiens – H. sapiens* PPI from each of the seven other species. It can be seen that the new normalization factor makes a slight difference in certain regions of the ROC curve. Figure 27 shows the ROC curves for predicting *M. musculus – M. musculus* PPI from the seven other species tested. The increase in accuracy at certain points in the ROC curve is much clearer for *M. musculus – M. musculus* prediction than *H. sapiens – H. sapiens* prediction. The ROC and P-R curves for other test species are included in the Appendix A (Figure 37 – Figure 52). Table 18 in Appendix A shows the performance metrics (AU-PRC and precision at 25% TPR) for each individual pair of training and testing species. A paired t-test was performed on the data in Table 18 to test if the true difference in the means of the performance metric for each normalization method was equal to zero. Table 7 shows the t-test results. The results show that the true mean of the differences between scoring methods is not equal to zero (p-value

< 0.0001 for both performance metrics). The true difference in mean precisions for the new normalization method compared to the old is estimated to be 1.1% and 1.9% for the AU-PRC and precision at 25% TPR metrics respectively (which represents a 6% and 9% increase in performance). Table 7 and Table 8 show that, while there are some test-species where it performs slightly worse, the new scoring method outperforms the old normalization method when using a single training species.

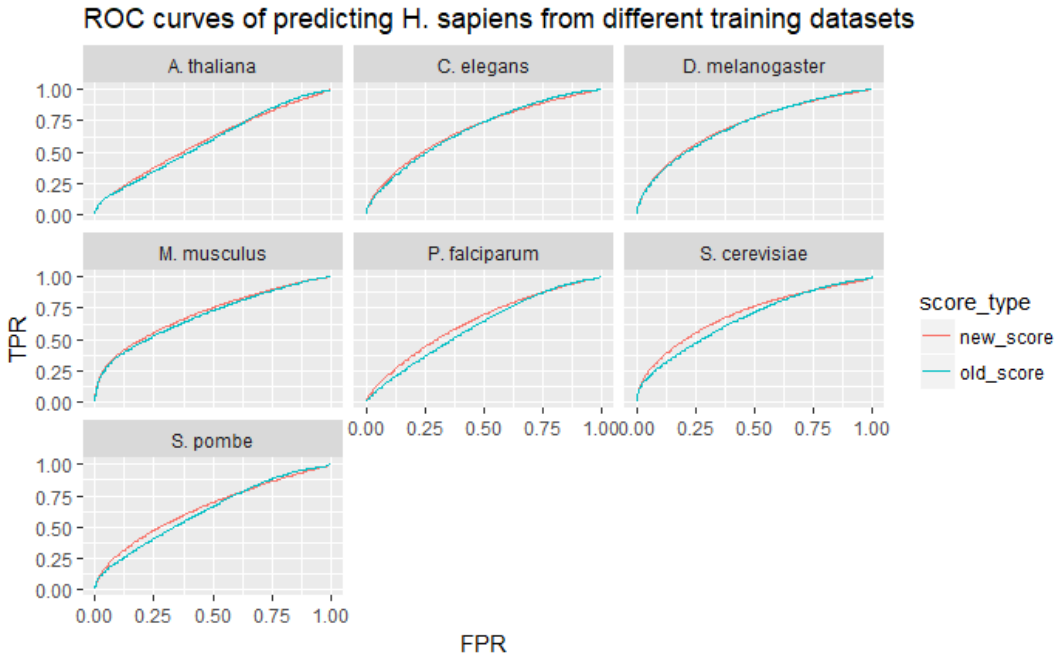



**Figure 26: Prediction of *H. sapiens* – *H. sapiens* PPI from seven training species. Subplot title gives the training species.**

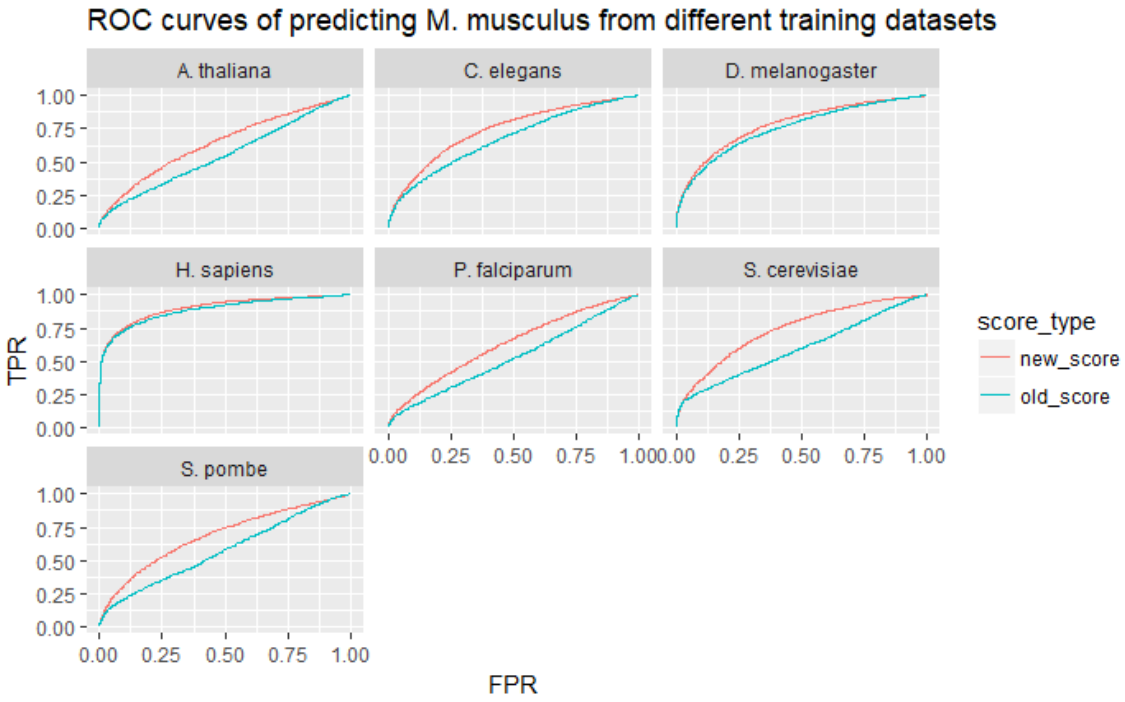



**Figure 27: Prediction of *M. musculus* – *M. musculus* PPI from seven training species Subplot title gives the training species.**

**Table 7: Paired t-test results comparing two methods of normalization for each performance metric for the single training species cross-species prediction.**

| | AU-PRC | Pr at 25% TPR |
|---|---|---|
| t-test difference in means (new – old) | 0.011 | 0.019 |
| t-test p-value | 2.87e-05 | 5.71e-05 |

**Table 8: Performance metrics for cross-species PPI prediction in different test species. Value is average score when predicting from each training species.**

| | AU-PRC | | Pr at 25% TPR | |
|---|---|---|---|---|
| **Test species** | **New** | **Old** | **New** | **Old** |
| *A. thaliana* | 0.163 | 0.148 | 0.164 | 0.142 |
| *C. elegans* | 0.190 | 0.193 | 0.206 | 0.209 |
| *D. melanogaster* | 0.157 | 0.142 | 0.166 | 0.141 |
| *H. sapiens* | 0.239 | 0.221 | 0.284 | 0.247 |
| *M. musculus* | 0.337 | 0.295 | 0.422 | 0.357 |
| *P. falciparum* | 0.127 | 0.110 | 0.130 | 0.111 |
| *S. cerevisiae* | 0.200 | 0.210 | 0.225 | 0.237 |
| *S. pombe* | 0.244 | 0.248 | 0.288 | 0.292 |
| **Average** | 0.207 | 0.196 | 0.236 | 0.217 |

#### 4.3.1.2 Multiple Cross-Species

Figure 28 shows the ROC curve for predicting *H. sapiens – H. sapiens* PPI from using the seven other species at once. It can be seen that the new scoring method performs significantly better than the original scoring method in this case. The performance change is more significant than the single cross-species prediction, as was expected. Figure 29 shows the ROC curve for each testing species when known interactions in the remaining seven species were pooled together as training data. Figure 30 shows the equivalent PRC. In Figure 30, there appear to be rapid changes in the precision at low TPR when predicting *A. thaliana* and *P. falciparum*. This is due the fact that there are very few predicted pairs at these extreme decision thresholds, and a single FP can significantly alter the precision and lead to such spikes. Table 9 shows the performance metrics for the two scoring methods for each test species. Table 10 shows the results of a paired t-test being performed on Table 9. The t-test results show that the true mean of the differences between scoring methods is not equal to zero (p-value $< 0.01$ for both performance metrics). The true difference in mean precision for the new normalization method compared to the old is estimated to be 9.6% and 16.4% for the AU-PRC and precision at 25% TPR metrics respectively (which represents a 45% and 67% increase in performance).

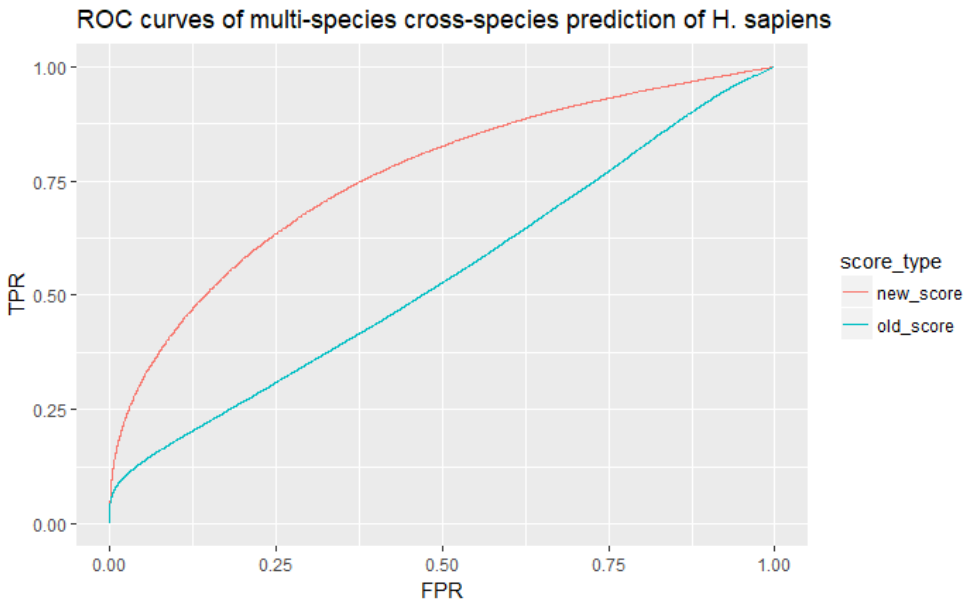



**Figure 28: ROC curve for multi-species cross-species prediction of *H. sapiens* PPI showing the performance of the two methods of normalization.**

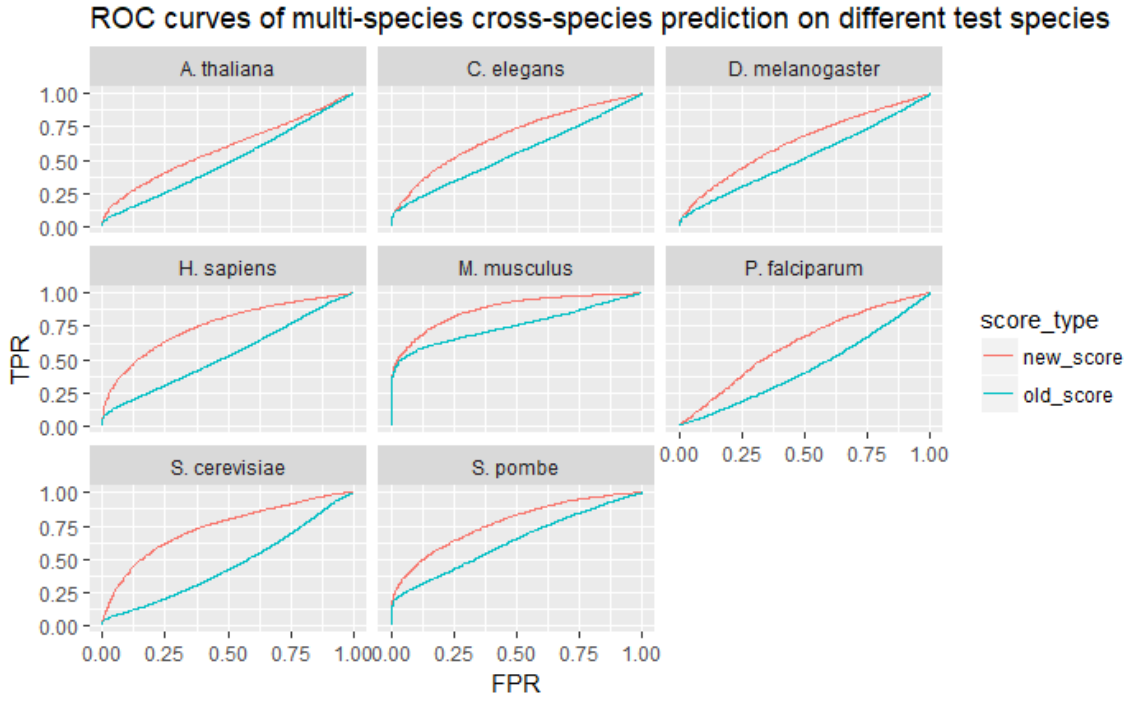



**Figure 29: ROC curves of using seven training species to predict intra-species PPI in the eighth species. Subplot title indicates test species.**

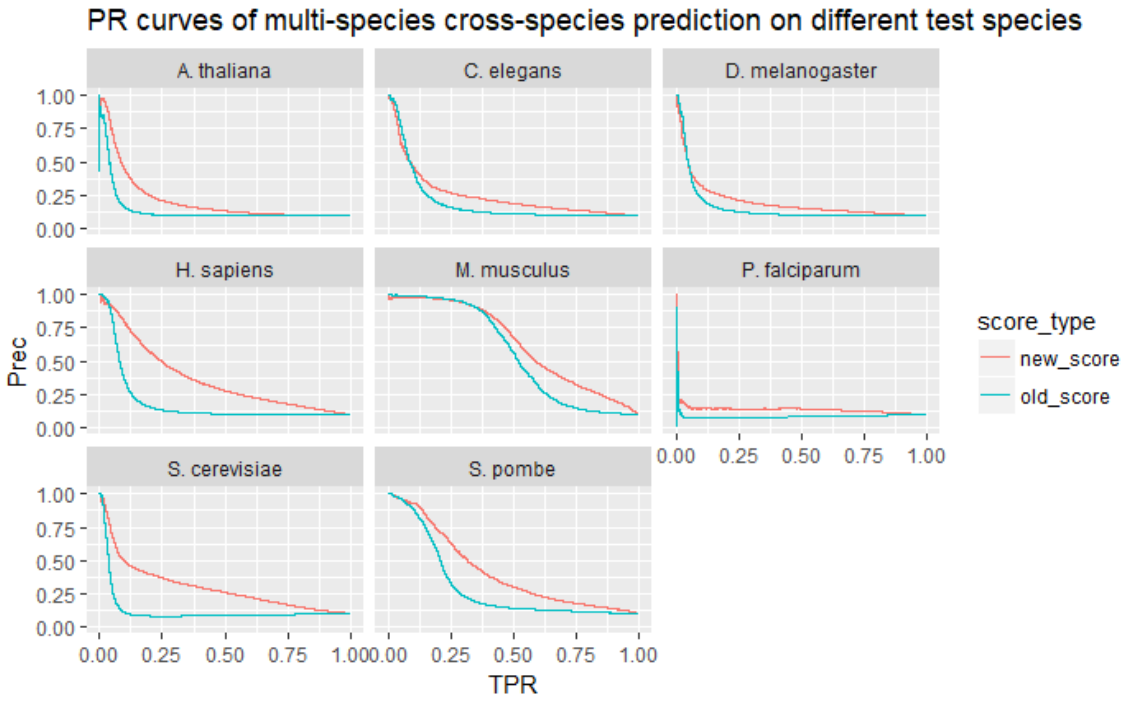



**Figure 30: P-R curves at 10:1 CI for each of the testing species when using the seven other species for training data. Subplot title indicates test species.**

**Table 9: Cross-species performance metrics when using multiple training species.**

| Test species | AU-PRC | | Precision at 25% TPR | |
|---|---|---|---|---|
| | New | Old | New | Old |
| *A. thaliana* | 0.214 | 0.138 | 0.211 | 0.104 |
| *C. elegans* | 0.249 | 0.197 | 0.269 | 0.160 |
| *D. melanogaster* | 0.200 | 0.156 | 0.213 | 0.127 |
| *H. sapiens* | 0.367 | 0.186 | 0.494 | 0.134 |
| *M. musculus* | 0.633 | 0.557 | 0.959 | 0.963 |
| *P. falciparum* | 0.140 | 0.090 | 0.144 | 0.078 |
| *S. cerevisiae* | 0.299 | 0.125 | 0.370 | 0.083 |
| *S. pombe* | 0.416 | 0.298 | 0.633 | 0.330 |
| **Average** | **0.315** | **0.218** | **0.412** | **0.247** |

**Table 10: Paired t-test results comparing two methods of normalization for each performance metric for the multiple training species cross-species prediction.**

| | AU-PRC | Pr at 25% TPR |
|---|---|---|
| t-test difference in means (new – old) | 0.096 | 0.164 |
| t-test p-value | 1.68e-03 | 9.94e-03 |

### 4.3.2 Evolutionary Distance Relation to Accuracy

Having established that the new score normalization approach is beneficial when using multiple training species, we next explore whether evolutionary distance can inform the selection of which training species to include for a given test species. Figure 31 shows average P-R curves resulting from predicting *H. sapiens – H. sapiens* PPI with a random sample of 2000 interactions (repeated twenty times) from each of the eight species. The P-R curves for the other species are given in Appendix A (Figure 53 – Figure 60). In each of the P-R curves, certain training species are far more effective than others when predicting intra-species PPI. In particular, when the training and test species are the same, maximum accuracy is observed. Figure 31 also appears to suggest that species that are more closely related to the test species may result in higher prediction accuracy (here, *M. musculus* is most closely related to *H. sapiens* and has the highest cross-species prediction accuracy). Table 11 and Table 12 show the performance metrics rankings for each training species when predicting *H. sapiens* interactions. Table 19 in Appendix A shows the performance metric rankings for each testing species.

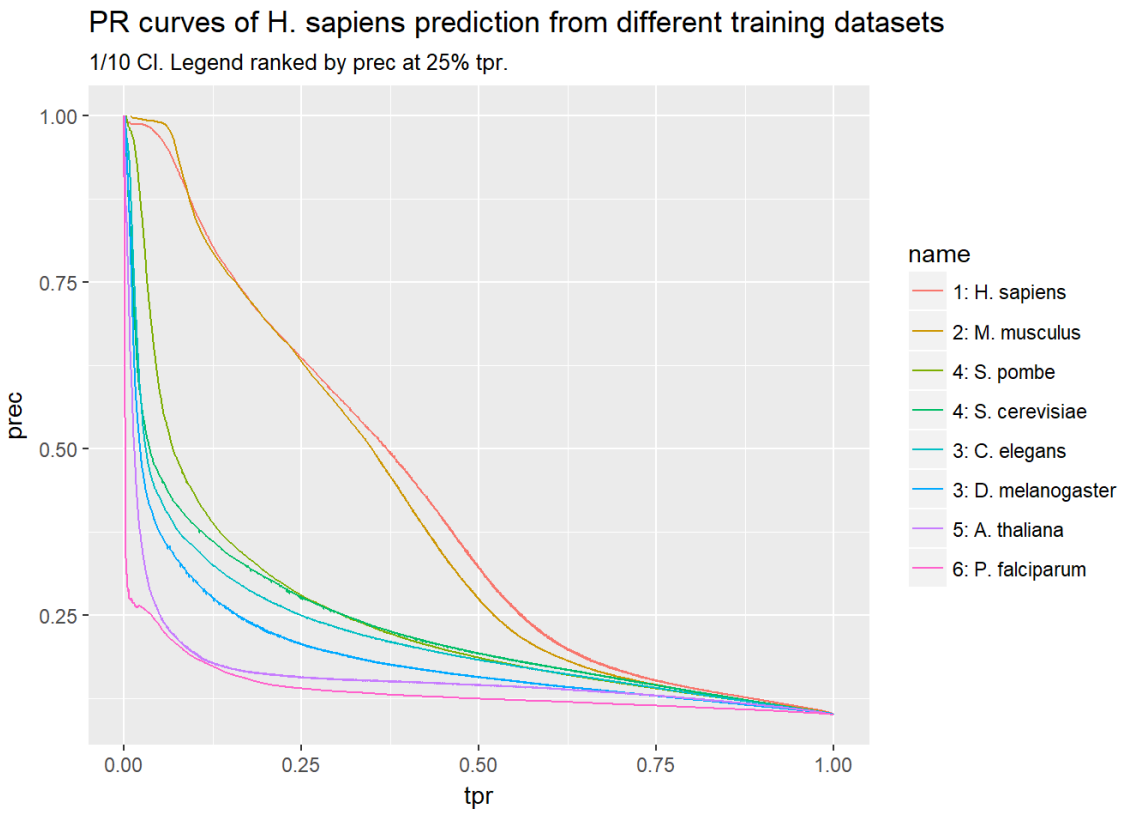



**Figure 31: P-R curve of using different species to predict known *H. sapiens* interactome. The data for each training species comprised 2000 randomly sampled interactions to control for interactome size. This was done 20 times and the performanc curves were averaged. Each line represents a different training species.**

**Table 11: Ranks and values of AU-PRC for predicting *H. sapiens* from other species. Here, rank_evo reflects relative distance (by number of common ancestors) between the training species and *H. sapiens*.**

| Species | AU-PRC | Rank_ AU-PRC | Rank_ Evo |
|---|---|---|---|
| *H. sapiens* | 0.409 | 1 | 1 |
| *M. musculus* | 0.388 | 2 | 2 |
| *S. pombe* | 0.241 | 3 | 4 |
| *S. cerevisiae* | 0.230 | 4 | 4 |
| *C. elegans* | 0.217 | 5 | 3 |
| *D. melanogaster* | 0.189 | 6 | 3 |
| *A. thaliana* | 0.159 | 7 | 5 |
| *P. falciparum* | 0.137 | 8 | 6 |

**Table 12: Ranks and values of Precision at 25% TPR for predicting *H. sapiens* from other species. Here, rank_evo reflects relative distance (by number of common ancestors), between the training species and *H. sapiens*.**

| Species | Pr@25%TPR | Rank_ Pr@25%TPR | Rank_ Evo |
|---|---|---|---|
| *H. sapiens* | 0.635 | 1 | 1 |
| *M. musculus* | 0.630 | 2 | 2 |
| *S. pombe* | 0.280 | 3 | 4 |
| *S. cerevisiae* | 0.277 | 4 | 4 |
| *C. elegans* | 0.249 | 5 | 3 |
| *D. melanogaster* | 0.206 | 6 | 3 |
| *A. thaliana* | 0.157 | 7 | 5 |
| *P. falciparum* | 0.140 | 8 | 6 |

At a very high-level, it appears that the evolutionary ranks of the training species when sorted by the performance metrics are proportional to the expected rank ordering from evolutionary distance, at least for *H. sapiens*. Table 13 shows the correlation coefficient between the expected ranks and the actual ranks of the AU-PRC performance metric on every test species for both the Kendall's Tau-b and Spearman correlation coefficient. The p-values are also found, through both classic statistical testing and through

permutation tests (100,000 permutations tested). The p-values with $p < 0.05$ are highlighted in bold. 16 out of the 32 p-values meet this significance criterion. Table 14 shows the equivalent table for the Pr@25%TPR performance metric. Here, 19 of 32 p-values meet the significant criterion for this metric.

**Table 13: Statistical tests for rank correlation between the expected rank (based on evolutionary distance) and the actual rank of the AU-PRC at 10:1 CI for each test-species. Permutation tests were done with N = 100,000. P-values smaller than 0.05 are shown in bold.**

| | Kendall's Tau-b correlation | | | Spearman's rank correlation | | |
|---|---|---|---|---|---|---|
| | corr. coeff. | p-value (from R) | p-value (permutation test) | corr. coeff. | p-value (from R) | p-value (permutation test) |
| *A. thaliana* | 0.68 | **0.033** | **0.018** | 0.76 | **0.027** | **0.017** |
| *C. elegans* | 0.44 | 0.132 | 0.087 | 0.52 | 0.188 | 0.099 |
| *D. melanogaster* | 0.82 | **0.006** | **0.002** | 0.93 | **0.001** | **0.001** |
| *H. Sapiens* | 0.67 | **0.024** | **0.014** | 0.80 | **0.018** | **0.012** |
| *M. musculus* | 0.82 | **0.006** | **0.002** | 0.89 | **0.003** | **0.003** |
| *P. falciparum* | 0.50 | 0.127 | 0.124 | 0.58 | 0.134 | 0.125 |
| *S. cerevisiae* | 0.32 | 0.288 | 0.184 | 0.34 | 0.406 | 0.206 |
| *S. pombe* | 0.48 | 0.111 | 0.076 | 0.58 | 0.129 | 0.069 |

**Table 14: Statistical tests for rank correlation between the expected rank (based on evolutionary distance) and the actual rank of the precision at 25% TPR at 10:1 CI (CI does not affect order for this metric). Permutation tests were done with N = 100,000. P-values smaller than 0.05 are bolded.**

| | Kendall's Tau-b correlation | | | Spearman's rank correlate ion | | |
|---|---|---|---|---|---|---|
| | corr. coeff. | p-value (from R) | p-value (permutation test) | corr. coeff. | p-value (from R) | p-value (permutation test) |
| *A. thaliana* | 0.68 | **0.033** | **0.018** | 0.76 | **0.027** | **0.018** |
| *C. elegans* | 0.52 | 0.079 | **0.049** | 0.58 | 0.133 | 0.070 |
| *D. melanogaster* | 0.74 | **0.012** | **0.006** | 0.88 | **0.004** | **0.003** |
| *H. sapiens* | 0.67 | **0.024** | **0.015** | 0.80 | **0.018** | **0.012** |
| *M. musculus* | 0.82 | **0.006** | **0.002** | 0.89 | **0.003** | **0.002** |
| *P. falciparum* | 0.50 | 0.127 | 0.125 | 0.58 | 0.134 | 0.125 |
| *S. cerevisiae* | 0.40 | 0.184 | 0.121 | 0.41 | 0.318 | 0.165 |
| *S. pombe* | 0.56 | 0.063 | **0.041** | 0.65 | 0.083 | **0.042** |

While the majority of the statistical test results were found to be significant at a $p < 0.05$ level, there were a number of test species where this was not the case. In particular, *P. falciparum* and *S. cerevisiae* did not appear to show preference for training data from species that were more closely related evolutionarily. In the case of *P.* falciparum, this effect may be due to the fact that all seven other training species are equally distant from it. Training with *P. falciparum* produced the highest accuracy of all species, but little else could be determined beyond that. For the case of predicting *S. cerevisiae* PPI, this effect may be due to the distantly related species, *P. falciparum*, performing very well.

The results of the modified Tau-b test, which was created to analyze all results together, are shown in Table 15. The p-value found through permutation test for this metric was significant at a $p < 0.0001$ level, giving strong evidence that over all test species, evolutionary distance between testing and training species is inversely correlated to accuracy of the prediction. However, as shown above, there may be specific testing species where this is not the case.

**Table 15: Modified Tau-b test to combine all data sources into one experiment. N = 10,000 in permutation test to find p-value.**

| | Custom Tau-b corr. coeff. | Custom Tau-b p-value |
|---|---|---|
| AU-PRC | 0.601 | 9.99e-05 |
| Prec at 25% TPR | 0.601 | 9.99e-05 |

## 4.4 Discussion

This chapter introduced a modified PIPE SW-score normalization function, specific to cross-species PPI prediction. When using a single training species to predict a test species, the modified scoring function produced superior results for most species tested. Although there were a few species where it performed slightly worse than the original scoring function, the increase in performance on most species was larger than the decrease

in performance on other species. A paired t-test was performed with the performance metrics for every pair of training and testing species; the t-test showed the new normalization factor was superior on both performance metrics, with a p-value < 0.0001. The true difference in mean precision is estimated to have improved by 6% and 9% for the AU-PRC and precision at 25% TPR metrics performance metrics respectively (10:1 class imbalance). As the number of training species increases, the effect of the normalization change becomes even more pronounced, as was initially hypothesized. When using seven testing species to predict an eighth, the AU-PRC (at 10:1 class imbalance) increased from 0.218 to 0.315 (i.e. a 45% improvement) with the new normalization method. Similarly, the average precision at 25% TPR increased from 0.247 to 0.412 (i.e. a 67% improvement). The paired t-test showed the new normalization factor was superior on both performance metrics, with a p-value < 0.01. As such, the new normalization factor can be recommended over the old factor for every type of cross-species PPI prediction.

The effect of evolutionary distance between training and testing species was also examined with PIPE. It was shown that classification performance was related to evolutionary distance between the training and testing species when controlling for training set size, with closer species performing better ($p < 0.0001$ with a modified Kendall's Tau-b test). However, there were certain test species that did not follow this pattern. This may indicate that the types of interactions known for each species may make different species uniquely suited for predicting other species. This may be due to the fact that known PPI for a given species often reflect experimenter's bias, where different types or families of proteins have been investigated to different degrees in different species, depending on the interest of the research community. Qualitatively, it appears that very closely related

species perform well in cross-species predictions, and that there is a steep drop-off in accuracy after this point. However, the size of the training set effect was not examined in detail; it may be that evolutionary distant species could still be useful once there is a sufficient number of known interactions.

Because of the change in normalization factor given in Section 4.2.2, it possible to generate the scores for protein pairs in a test species using each training species separately and individually. The scores from each training species for a specific protein pair can then be combined by simple summation. If there are known interactions in the test species, this change allows us to easily examine different combinations of training species for each test species to see if there is a way to use all data to produce better results. This would have been difficult previously, as PIPE would need to be run for every combination of training species. Additionally, PIPE was shown to significantly decrease in performance when using multiple training species without the normalization factor change. In our preliminary experiments, no systematic rule was found regarding what combination of species should be used for training in a test species. In some test species, using a combination of species performed better than any single species. However in other test species, the best performance occurred when using a single, closely related species. In the future, general rules should be developed regarding what combinations of test species should be used for predicting test species without known interactions on which to test.

## 4.5 Conclusion

In this chapter, the use of PIPE in cross-species prediction was explored and improved. A new scoring method was created for PIPE, specifically for to cross-species PPI prediction use case. The new scoring method was shown to be superior to the original

scoring method, especially as the number of training species increases. It was then shown that cross-species PPI prediction performance was related to the evolutionary distance between the training and testing species when controlling for training set size. However, the size of the training set effect was not examined in detail.

When developing a PPI predictor for a new test species, it is recommended that several subsets of training species be considered. With the new scoring method, rapid evaluation of different training species subsets is possible, since each training species may be considered in isolation and combined through simple summation after the fact. If there are no known interactions in the test species this such optimization is not possible. In this case, it is difficult to provide recommendations on which training species to use. It is recommended that, if there exist a closely related species with many known interactions, then that species should be used alone for training. Otherwise, a combination of several, potentially distant training species could be used. As more training data becomes available in the future, then using data from multiple species could become even more useful in making new predictions. These findings should be re-examined in the future as more complete interactomes become available through high-throughput experimental techniques.

# 5 Inter-species PPI prediction

## 5.1 Introduction

Inter-species PPI prediction refers to predicting interactions between proteins arising from different species. The most common example of inter-species PPI would be between a pathogen and its host. Other examples include the allergic response between a host and a food (e.g. human allergy to soy proteins) and PPI between symbiotic organisms. This chapter will examine the suitability of PIPE to predict inter-species interactions, which had not been done previously. To this end, PIPE is applied to the prediction of inter-species PPI between the Human Immunodeficiency Virus-1 (HIV-1) and human. These species were chosen since they are both relatively well-studied and a number of human – HIV-1 PPI have been documented and are available to evaluate the inter-species predictions. The HIV-1-Human PPI prediction sections of this chapter appeared previously in a conference paper, for which I was the primary author [71].

Once it is established that PIPE is able to make inter-species predictions, PIPE is then applied to make inter-species predictions using intra-species PPI from different species – an example of both inter-species and cross-species PPI prediction. This will be done for the motivating example of the thesis: predicting novel PPI between the SCN pest and soybean, using cross-species training data from *A. thaliana* and *C. elegans*. These results are of interest to researchers at Agriculture and Agri-Food Canada, who are seeking to develop ways to prevent and reduce the effect of the SCN pest on soybean crops. In addition, researchers at Agriculture and Agri-Food Canada were also interested in

researching the soybean allergy in humans, so a final inter-species PPI prediction was completed between soybean and human, using *A. thaliana* and *H. sapiens* as training data.

## 5.2 Methods

### 5.2.1 HIV-1 – Human Prediction

In order to use the PIPE algorithm to predict interactions between viruses and humans, two different data sources are needed: a list of known interactions between virus and human proteins for training and validation, and the primary sequence data for each protein in the training and prediction data sets. The known interactions were taken from the HIV-1 Human Interaction Database [28]. This database consists of high-quality, manually-curated interactions between HIV-1 and human proteins. The interactions database was filtered to only include the 'binds' interaction type, since the database also documents other irrelevant interaction types, such as “genetic” interactions where one protein affects the expression levels of another protein. Duplicate interactions were removed by comparing the exact sequences downloaded from RefSeq [72]. All human sequence identifiers were mapped to UniProt Sequence IDs (Swiss-Prot where possible) using sequence identity and UniProt's ID/Mapping tool [68]. To obtain a complete human proteome (i.e. all human proteins, not only those appearing in known PPI), the Swiss-Prot human proteome was used [68]. The HIV-1 proteome consisted of all 18 HIV-1 proteins; sequences were downloaded from RefSeq [72]. The sizes of the various proteomes and interactomes are summarized in Table 16 below.

**Table 16: Data summary of the HIV-1 Interactions Database and Swiss-Prot human.**

| | |
|---|---|
| Unique HIV-1 – human interactions in HIV-1 Human Interaction Database | 7,567 |
| Number of 'binds' interactions (after cleaning) | 611 |
| Number of virus proteins (all interact) | 18 |
| Number of human proteins involved in interactions | 492 |
| Total number of human proteins | 20,231 |
| Number of HIV-1 – human protein pairs to examine | 363,547 |
| Largest number of known HIV-1 interactions for a single human protein | 6 |
| Largest number of known human interactions for a single virus protein | 215 |

To evaluate the performance of PIPE, leave-one-out cross-validation (LOOCV) was performed. In this procedure, the classifier was trained using the filtered interactions database minus one interaction; PIPE was then used to generate a score for the interaction. This process was repeated for each interaction in the filtered interactions database. PIPE was then used to generate interaction scores for a negative data set, which consisted of randomly generated human – HIV-1 protein pairs that were not known to interact. This approach may produce false negatives, but the number will be low due to the rarity of PPI among all possible protein pairs. With the PIPE scores calculated for the negative and (leave-one-out) positive data sets, the ROC and P-R curves were generated.

Since the SW score and the traditional PIPE score (which uses a filter on the interaction landscape) have not been compared for inter-species PPI prediction, they will both be generated for the LOOCV testing data. Note that the original formulation of the SW score (without the changes in Section 4.2.2) was used. This was because HIV-1 has a total of 18 proteins; normalizing by how many of the 18 proteins a window was seen in would be likely to bias the score unnecessarily. After the ROC curves for both scoring methods have been generated with LOOCV, the superior scoring method will be used to

compare PIPE to other state-of-the-art HIV-1 – human PPI prediction methods outlined in Section 2.3.2.

### 5.2.2 SCN – Soybean PPI Prediction

The case of inter-species PPI prediction between SCN and soybean is an example of both inter-species and cross-species PPI prediction. The actual predicted PPI will be inter-species PPI. However, since there are no known interactions between SCN and soybean, and very few are known intra-species PPI in either SCN – SCN or soybean – soybean, we will have to use cross-species prediction by using training data from a similar species for both soybean and SCN. Based on the results in Chapter 4 for cross-species prediction, we decided to use the known interactions from two model species that are closely related evolutionarily to each soybean and SCN. These species are *A. thaliana* for soybean and *C. elegans* for SCN. Only these two training species were used, as they are both well-studied and closely related to the testing species. Other possible training species are significantly more distant evolutionarily. For example, *H. sapiens* would be the next most similar species after *A. thaliana* for soybean prediction (as seen in Figure 25 above). Figure 32 shows a high-level overview of the data used to make predictions between soybean and SCN. The known interactions from *A. thaliana* and *C. elegans* were pooled together, as were the sequences from all four species, and PIPE was run on every pair of proteins between the 75,781 soybean proteins and the 21,868 SCN proteins. The sequences for soybean and SCN were received directly from researchers at Agriculture and Agri-Food Canada. The sequences and interactions from *C. elegans* and *A. thaliana* were found via a similar process to the one outlined in Section 4.2.1. The only difference was that all TrEMBL sequences were used in addition to the Swiss-Prot sequences for *C. elegans*, since

there were fewer Swiss-Prot sequences. The new normalization factor outlined in Section 4.2.2 was used to calculate the PIPE score.

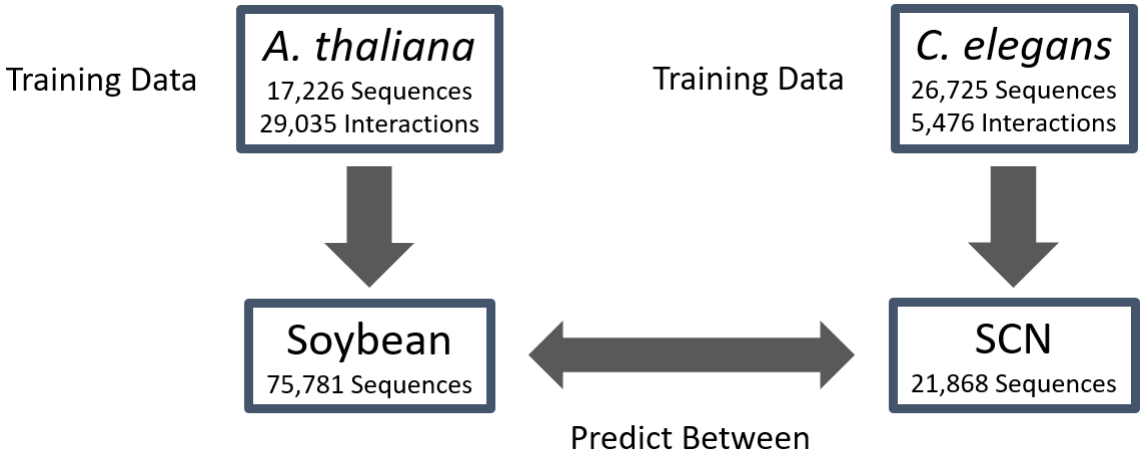



**Figure 32: High level overview of how training data were used to predict inter-species PPI between soybean and SCN using the known interactions from two similar model species. Each model species chosen was closely related to one of the species of interest.**

### 5.2.3 Human – Soybean PPI Prediction

For the prediction of inter-species interactions between human and soybean, a similar setup was used, as seen in Figure 33. Since human has many known intra-species interactions, the only other model species necessary was *A. thaliana*. All known interactions and sequences from the three species were pooled together, and PIPE was run on every pair of proteins between the 75,781 soybean proteins and the 20,236 human proteins. The sequences for soybean and SCN were received directly from researchers at Agriculture and Agri-Food Canada. The sequences and interactions from *H. sapiens* and *A. thaliana* were found by the process outlined in Section 4.2.1.

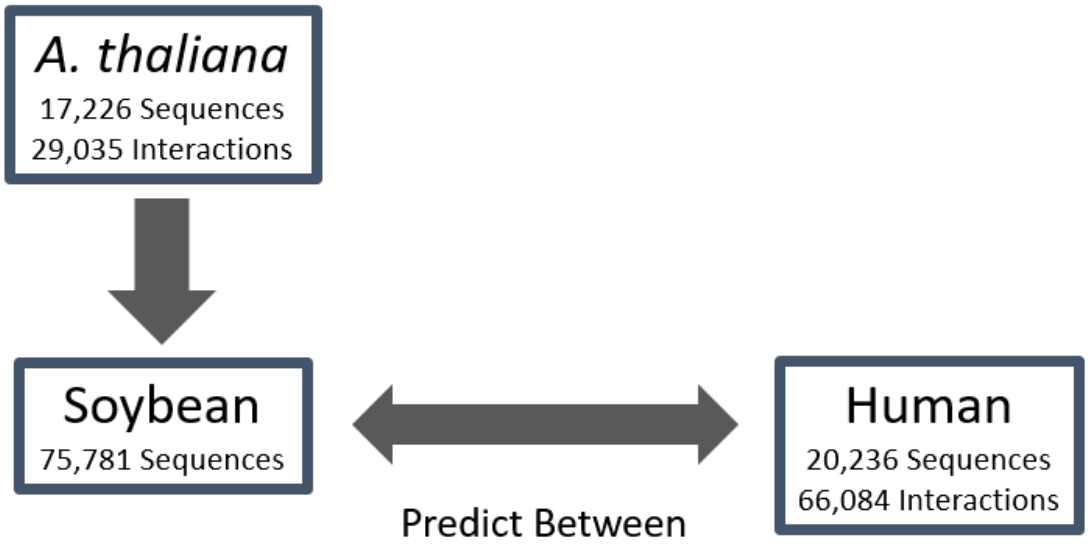



**Figure 33: High-level overview of the data used to prediction inter-species PPI between soybean and human.**

## 5.3 Results

### 5.3.1 HIV-1 – Human PPI Prediction

In order to analyze the performance of using PIPE to predict novel HIV-1 – human PPI, LOOCV was performed. The receiver operating characteristic (ROC) curve, shown in Figure 34 below, shows the FPR and TPR of the classifier for different score cut-off thresholds. As the SW score is superior to the traditional PIPE score, especially in the region of low FPR, it was used for further plots and predictions.

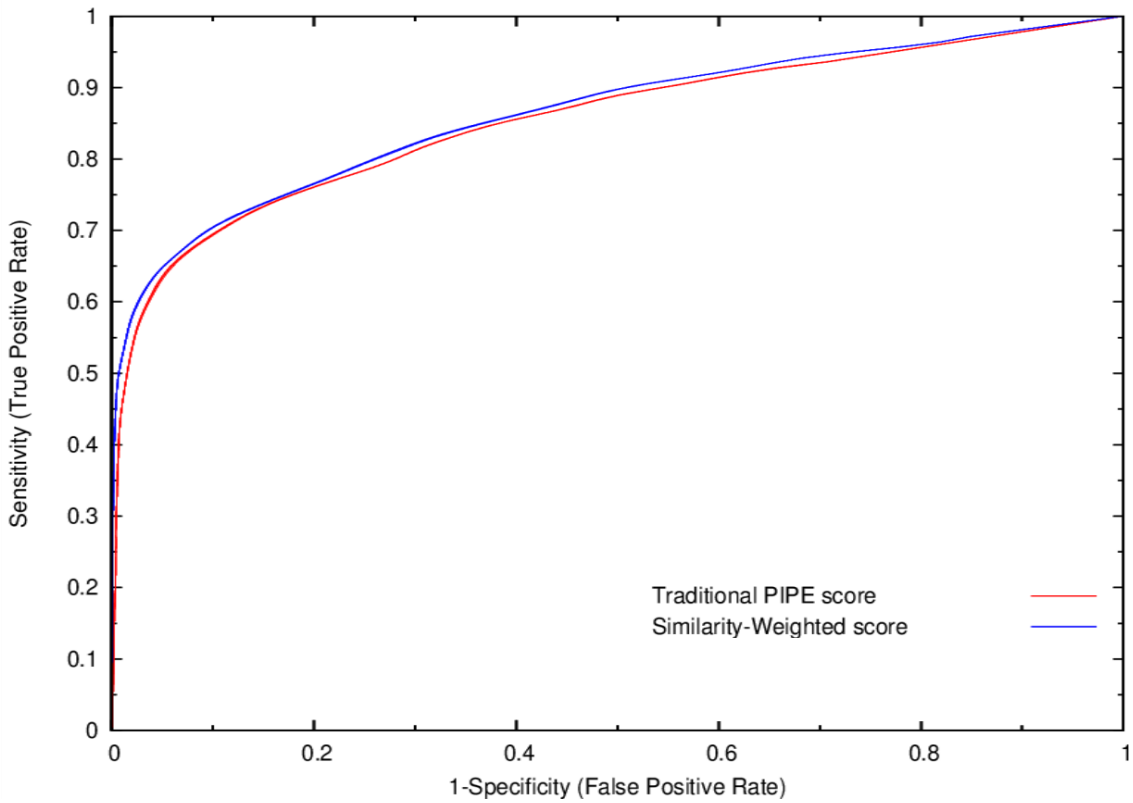



**Figure 34: Receiver operating characteristic (ROC) curve for HIV-1 – human PPI prediction. ROC curve was found through LOOCV using PIPE to predict interactions in the HIV-1 human interaction database.**

The ROC curve does not entirely illustrate how the classifier will perform on actual data, as it does not account for the class imbalance. Since there are already 611 known binding interactions out of a possible 363,547 protein pairs (Table 16), the class imbalance can be no greater than 600:1. Previous studies have used a class imbalance of 100:1 [48], [51], [47]. The performance of the algorithm for different class imbalances is shown as P-R curves in Figure 35. Higher class imbalance essentially makes the problem more difficult, leading to lower achievable precision for the same recall level. Due to the class imbalance, the classifier should operate at a high specificity, or the number of false positives will overwhelm the number of true positives predicted, decreasing the precision.

In our case, we selected a cut-off value corresponding to a specificity of 99.95% and a recall of 22.5%, given by the black vertical line in Figure 35.

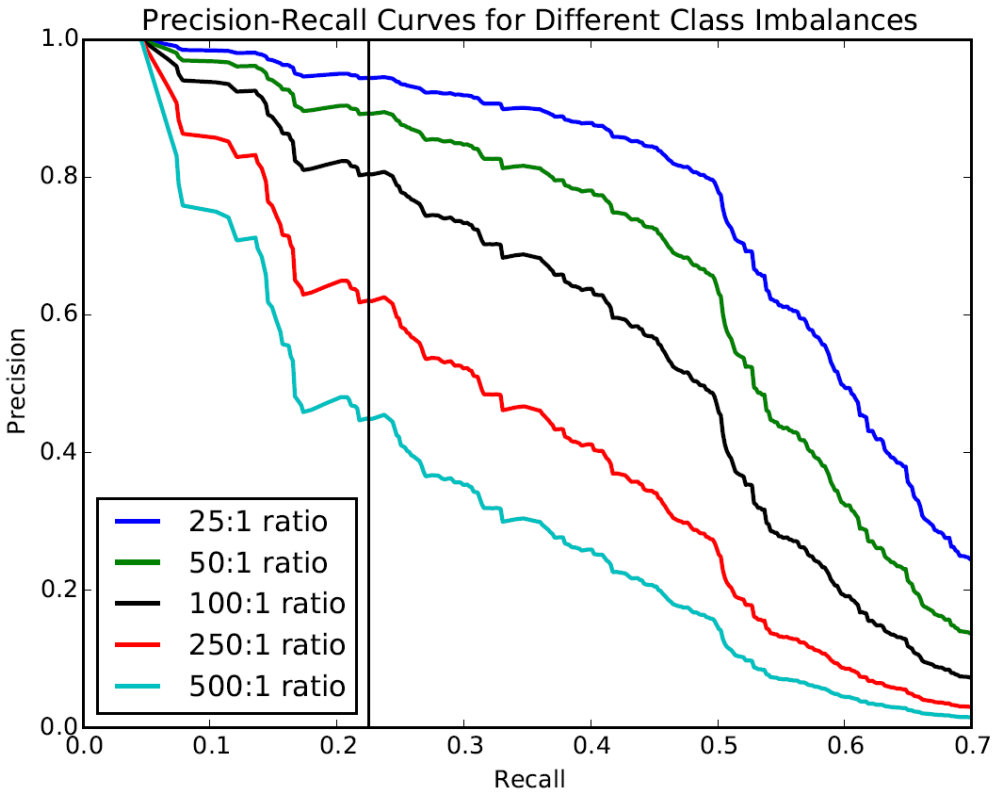



**Figure 35: Precision-recall curve for HIV-1 – human PPI prediction. The precision is shown for different class imbalance levels. The selected decision threshold, shown by the vertical black line, resulted in a recall of 22.5%.**

The cut-off score for determining whether a protein pair interacts can be set to different levels depending on the number of novel interactions desired, the cost to experimentally verify an interaction, and the biological relevance of the potential interaction. The predicted protein pairs can also be examined in order of PIPE score, which is directly correlated to the expected precision for the protein pair. For different recall levels (and corresponding PIPE score cut-offs), the number of novel interactions and the number that are expected to be valid are shown in Figure 36. A class imbalance ratio of 100:1

negative to positive interactions was used alongside the given recall to determine the precision, and thereby the number of expected valid interactions.

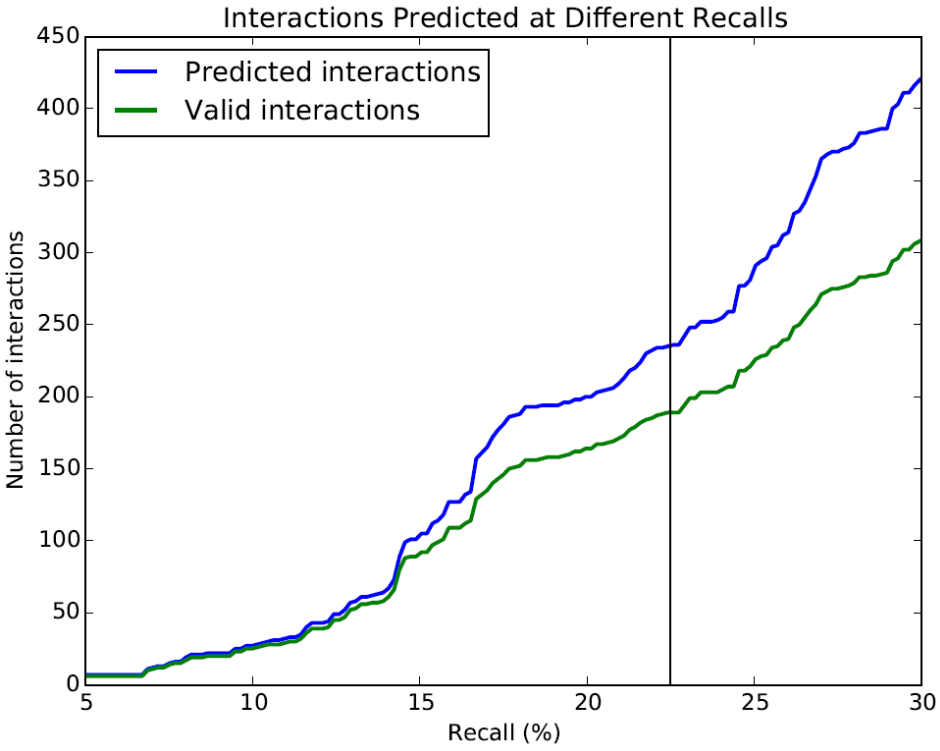



**Figure 36: Plot showing total number of predicted interactions and the number that are expected to be true interactions as a function of recall. The difference between the two curves represents the number of anticipated false positives. A class imbalance ratio of 100:1 was used to determine the precision and thereby the number of expected valid interactions. The decision threshold was set at a recall of 22.5%, resulting in 229 novel predictions, with 188 of them expected to be valid.**

A highly conservative cut-off value was selected, corresponding to a recall of 22.5% with a specificity of 99.95%. This threshold resulted in 229 novel interactions predicted. The precision of these predictions would depend on the class imbalance. At a conservative estimate of class imbalance (500:1), the precision is 47.4%. The precision for other possible class imbalances are shown in Table 17.

**Table 17: Precision estimate for the 229 novel PPI predictions at different class imbalances at the final decision threshold.**

| Class Imbalance | 500:1 | 250:1 | 100:1 | 50:1 | 25:1 |
|---|---|---|---|---|---|
| Precision | 47.4% | 64.4% | 82.0% | 90.2% | 94.9% |

### 5.3.2 SCN – Soybean and Human – Soybean PPI Prediction

With the improved algorithm outlined in Chapter 3, and the modified SW Score normalization outlined in Chapter 4, all 1,657,178,908 pairs of proteins between SCN and soybean were run through PIPE on a local cluster over a period of several days. Similarly, all 1,594,212,316 pairs of proteins between human and soybean were run also run over a period of several days. The actual runtimes can be seen in Table 4 in Section 3.5. These results were sorted by SW score, providing a ranked list of probable interactions to examine. They were also indexed by protein, so that researchers could examine the most likely interactions involving any protein of interest. These results were provided to Agriculture and Agri-Food Canada on an external hard drive. The results files for all scores totaled over 500 GB. Since there are no known interactions between SCN and soybean, it is difficult to assess the accuracy of the predictions to traditional pattern classification schemes.

## 5.4 Discussion

### 5.4.1 HIV-1 – Human

The results in Table 17 are strong results for the given class imbalances, and they show that PIPE matches or exceeds the performance of previous HIV-1 human PPI prediction methods. The Dyer method developed by Dyer *et al.* represents the state of the art for HIV-1-human PPI prediction [51]. Our performance is similar to that of Dyer;

however, when operating at very high specificity (required due to the class imbalance of 100:1 or higher), our method makes more positive predictions and achieves a higher precision. Using a class imbalance of 100:1, Dyer *et al.* predicted 189 novel interactions at a precision of 70%, while PIPE predicted 229 novel interactions at a precision of 82%. Tastan *et al.* [48] and Qi *et al.* [49] achieved a precision of 23% and 27.7% respectively at a class imbalance of 100:1. Even at the conservative estimate for class imbalance (500:1), PIPE achieves a precision of 47.4%, showing that PIPE matches or exceeds previous work for HIV-1 human PPI prediction. Since different PPI prediction methods often use slightly different training data and different features, the predicted PPI often do not overlap completely. These methods can therefore be seen as complementary rather than competitive. This chapter shows that PIPE can be used in addition to other state-of-the-art methods to improve the state of HIV-1 – human PPI prediction. A number of other studies have been published, but are not compared here due to methodological issues such as the use of unrealistic class imbalance, inadequate performance measures, or applicability to only a small fraction of protein pairs due to the requirement for known protein structure.

Due to the high specificity required to maintain a reasonable precision, the recall was generally limited to 22.5%; this results in the majority of true interactions being missed. However, the high precision would mean that wet-lab validation experiments are likely to produce positive results. Due to the high cost of such experiments, this is an important consideration.

One limitation of any machine learning-based method of PPI prediction, including PIPE, is that they may miss interactions that do not resemble any previously known interactions. This effect could also result in somewhat optimistic LOOCV performance

estimates. If there are many interactions between similar proteins in the training data, the algorithm would likely continue to predict each of these interactions if only one of them is removed during LOOCV. Conversely, randomly generating negative interactions may have resulted in the inclusion of false negatives during LOOCV; maintaining a high specificity despite these false negatives would potentially require an overly conservative decision threshold, thereby limiting the achievable recall.

### 5.4.2 Under-studied Viruses

During the Zika virus outbreak, PIPE was used to predict novel PPI between the Zika virus and human. Another sequence-based PPI prediction method for viruses, DeNovo [60], previously found that using virus-human interactions between closely related viruses resulted in the best accuracy when predicting PPI in under-studied viruses. PIPE was applied to Zika-Human, using known PPI data between human and other viruses of the family (*Flaviviridae*). DeNovo was also run to predict Zika – human interactions. A biological analysis of the resulting predictions was published in [73]. While I provided technical contributions to the published paper by running PIPE and preparing the data, no novel experiments were performed directly relating to this thesis topic.

### 5.4.3 SCN – Soybean

Using the best practices discovered in this thesis, we performed a global prediction between all SCN and soybean proteins as well as between all soybean and human proteins. As stated in Chapter 1, this was the motivating example of the thesis, and it leveraged aspects of all the contributions made in this thesis. Unfortunately, it is difficult to assess the accuracy of these predictions, as there are no known interactions between either SCN

and soybean or human and soybean. In the near future, the top scoring protein pairs will be examined from a biological standpoint to determine if the interacting proteins take part in biological processes that are expected to be affected by the other species. Preliminary feedback from biological researchers at Agriculture and Agri-Food Canada are pleased with the PIPE predictions as they have confirmed some initial hypotheses, which they plan to publish in the coming year.

## 5.5 Conclusion

This chapter showed that PIPE has the potential to predict inter-species PPI in addition to intra-species and cross-species PPI. PIPE was then used to predict all possible PPI in three different inter-species examples, including between HIV-1 and human, SCN and soybean, and human and soybean. Of these three inter-species examples, only the HIV-1 – human example had known positive interactions with which to test classification accuracy. It was shown that PIPE is comparable to the state of the art in HIV-1 – human PPI prediction, able to make predictions at a recall of 22.5% and specificity of 99.95%. In total, 229 high-confidence novel interactions were predicted. Depending on the estimated class imbalance, the precision for these predictions ranges from 47.4% to 90.1% (corresponding to class imbalances of 500:1 to 50:1). The PPI prediction results will be provided to HIV-1 researchers, with the hope that predicted novel interactions could lead to an improved understanding of the HIV-1 virus and potentially new treatment targets.

The other two inter-species PPI predictions were done at the behest of Agriculture and Agri-Food Canada, and they serve as the motivating examples behind all other work done in this thesis. Preliminary feedback from these collaborators, based on unpublished

data, has been positive. Large-scale validation of predicted PPI is planned in the near future.

# 6 Thesis Summary and Future Recommendations

## 6.1 Conclusions

Over the course of this work, a leading sequence-based PPI predictor, PIPE, was made more suitable for a new type of PPI prediction that is important to researchers at Agriculture and Agri-Food Canada. The new type of PPI prediction was a combination of inter-species and cross-species prediction, where PPI were predicted between two under-studied species with very few known interactions. The motivating example for this topic was the case of the Soybean Cyst Nematode (SCN), an extremely costly pest that causes billions of dollars of lost soybean crops every year. Discovering new PPI between SCN and soybean has the potential to help develop new resistant strains of soybean and other ways of reducing the damage of the SCN pest.

In tackling this problem, several contributions were made towards PIPE, especially as it relates to making new predictions in and between under-studied species. The first contribution was made through analyzing the algorithms behind PIPE: by using a larger and different type of database format, PIPE was made significantly faster. This allowed us to predict the 1.6 billion possible interactions between soybean and SCN on a local cluster in under two days, something which would have previously taken over forty days. PIPE was then made more accurate for cross-species PPI prediction involving multiple species by a modification to the scoring algorithm. The best datasets to use for cross-species prediction were then examined; it was found that using training species evolutionarily similar to the testing species would have the highest performance. PIPE was then shown to be capable of predicting inter-species interactions, such as between a pathogen and host.

PIPE was comparable to the state of the art for predicting interactions between HIV-1 and human.

Leveraging all of the contributions previously mentioned, the problem of predicting PPI between SCN and soybean was tackled. Using the improved algorithm speed, modified scoring function, and recommendations for best data sets, all possible pairs of SCN and soybean proteins were examined to predict novel interactions. These results were provided to researchers at Agriculture and Agri-Food Canada. While the exact performance metrics of this new prediction cannot yet be known, the researchers were pleased with the results as it helped confirm some preliminary hypotheses which they are unable to share at this time. Additionally, prediction of all human-soybean PPI was completed using the improved and accelerated version of PIPE. Agriculture and Agri-Food Canada researchers are looking into finding out the cause of the soybean allergy experienced by many people in order to develop non-allergenic soybean strains.

## 6.2 Summary of Contributions

In Chapter 3, the PIPE algorithm was made significantly faster by using a modified data structure and algorithm. For intra-species PPI prediction, PIPE was made 53x, 13x, and 8x faster for *H. sapiens*, *A. thaliana*, and *S. cerevisiae* respectively. Overall, the longer the PIPE algorithm would previously take per protein pair, the larger the speed improvement with the modified algorithm. The modified algorithm allowed the motivating example of an all-to-all SCN-soybean prediction to be completed on a local cluster of 18 nodes in 2 days instead of the 43 days it would have taken previously. With the time saved, the cluster was used to make an all-to-all prediction between human and soybean in 3 days instead of the 72 days it would previously have taken. The increased speed will make it

easier to apply PIPE to a variety of applications in the future. It will also aid InSiPS, which is a genetic algorithm that uses repeated PIPE runs to design new proteins that interact with specific proteins of interest [74]. There are certain downsides to the modified version of PIPE, including needing to re-run the pre-processing step if the known interactions list changes, the complication of LOOCV testing, and the increase in database file size. However, none of these drawbacks significantly impact the most common use-case of PIPE: large all-to-all predictions that previously took several weeks on a medium sized computer cluster.

In Chapter 4, the use of PIPE for cross-species prediction was examined in detail. It was shown that classification performance was related to evolutionary distance between the training and testing species when controlling for training set size, with closer species performing better ($p < 0.0001$ with a modified Kendall's Tau-b test). However, there were certain test species that did not follow this pattern. This may indicate that the types of interactions known for each species may make different species uniquely suited for predicting other species. Also, the size of the training set effect was not examined in detail; it may be that species that are more evolutionary distant could still be useful once there is a sufficient volume of known interactions. Chapter 4 also introduced a modified scoring function specific to cross-species PPI prediction. The new scoring method performs better for cross-species prediction, especially as the number of species involved increases. When using seven testing species to predict an eighth, the average area under the precision-recall curve (at 10:1 class imbalance) increased from 0.218 to 0.315 (a 45% increase). Similarly, the average precision at 25% sensitivity increases from 0.247 to 0.412 (a 67% increase). Even when using a single training species, the modified scoring function generally makes

a positive impact (however there were cases where it performed very slightly worse). An additional benefit of the modified scoring function is that training species can be used to predict PPI in test species individually, and then combined after the fact. If there are known interactions in the test species, it would be possible to test what combinations of training species produce the best results. This method has the potential to increase same-species PPI prediction as well if interactions from multiple species are used in addition to the species of interest.

In Chapter 5, we showed that PIPE is comparable to the state-of-the-art method for predicting interactions between Human Immunodeficiency Virus-1 (HIV-1) proteins and *H. sapiens* proteins, an example of inter-species PPI prediction. An all-to-all between HIV-1 proteins and *H. sapiens* proteins was performed, and the results were provided to HIV-1 researchers. Chapter 5 also outlines the prediction of SCN – soybean interactions. Using the best practices developed in the thesis, all possible SCN – soybean PPI were predicted using the accelerated PIPE implementation, the modified scoring algorithm, and training data from two model species evolutionarily close to SCN and soybean (*C. elegans* and *A. thaliana* respectively). Afterwards, a prediction between all soybean and human proteins was completed as well. All results were provided to our collaborators at Agriculture and Agri-Food Canada. While our collaborators are pleased with the results, no performance metrics are available for these results since ground truth PPI data is not yet available.

## 6.3 Recommendations for Future Work

### 6.3.1 Computational Improvement

During my thesis research, the SPRINT algorithm was developed [38]. Analysis of their implementation has suggested a further improvement to PIPE. Instead of creating a Type 2 DB file as outlined in Figure 20 in Section 3.3, PIPE could be run with only the Type 1 DB file. This can be accomplished by modifying the new PIPE algorithm (outlined in Figure 22) to loop through known interacting pairs directly instead of individual proteins. The Type 1 DB file could then be used for each similar window to increment the interaction landscape. While the Big O runtime would be the same, this modification would have several benefits, including enabling LOOCV, preventing the need for re-running the pre-processing step if the known interactions list changes, and reducing the amount of space required to store the database files. Another PIPE change that could be made would be to stop storing the complete interaction landscape, and just store the score when calculating it with PIPE. This would significantly reduce the memory required by PIPE, allowing more threads to be run at once and further increasing speed. While the change would not affect the relative speed of the algorithm changes made in this thesis, it could be used to increase the speed of both algorithms by increasing the number of threads available to run PIPE. Additionally, the memory usage saved could be used to store the similar proteins database files in memory instead of on a solid-state drive, which could potentially speed PIPE up even further (this effect would be even more impactful if PIPE was run on computing clusters with regular hard disk drives, which have significantly slower read times than solid-state drives).

It was shown that the new implementation of PIPE was significantly faster than the original implementation of PIPE. However, the exact speed increase is not known for each use-case (rather an approximate speedup value is given in this thesis). The exact speed increase should be further verified by repeating the timing experiments multiple times, for different species, and on different computing clusters.

The increasing speed of PPI prediction methods could have broader implications for the field at large. First, it will become easier to test different scoring systems, such as the normalization change examined in Chapter 4. Previously, the computational cost of examining many different scoring methods and parameters made it difficult to rigorously optimize these scoring functions. Second, the increased speed will make it possible to regularly re-run and update whole-interactome predictions as more interactions become known. The new predictions could be used to guide biological experiments, and the results could then be supplied back into the PPI predictor to produce even better predictions. Third, online databases that collect PPI predictions for summarization and biological analysis could be extended to include high-confidence predicted interactions for multiple species, so that biologists can easily examine predicted interactomes for any species of interest.

### 6.3.2 Cross- and Inter- Species PPI Prediction

With the inter-species and cross-species PPI prediction setup improvements in this chapter, SCN-soybean and human-soybean PPI were able to be predicted. It is hoped that the findings of this thesis may be applied in the future to other cross- and inter-species PPI predictions, of interest to Agriculture and Agri-Food Canada as well as general public health researchers. Future cross-species PPI predictions could be useful to researchers seeking to understand and modify other crops including soybean, rice, and corn. Inter-

species prediction could be applied to other pathogens affecting humans, including Hepatitis C, malaria, and tuberculosis. PIPE could be also applied to other pests affecting crops and even livestock.

The predicted interactions between HIV-1 and human should be examined by HIV researchers in the future. Interactions that involve important biological processes in HIV-1 infection may be further examined and verified experimentally. The PIPE algorithm requires only sequence data, but it still matches the performance of algorithms that use a variety of other data sources, such as structure, expression levels, or gene ontology process. Combining PIPE score with these other data sources has the potential to significantly improve the state of the art for predicting HIV-1 human PPI.

This thesis showed that PIPE was suitable for cross-species and inter-species predictions. However, no experiments were completed that involved using intra-species interactions (e.g. known human-human PPI) to predict inter-species interactions (e.g. HIV-1 – human PPI) when there were known inter-species PPI to test performance. It would be useful to show the accuracy of using intra-species PPI instead of inter-species PPI to predict inter-species PPI. The results of this experiment could potentially be used to improve general inter-species PPI prediction if using known intra-species PPI in addition to known inter-species PPI produced superior results to either one on its own. Using intra-species PPI to predict inter-species PPI could also make it possible to more accurately predict PPI between human and other under-studied viruses by using known human-human interactions. HIV-1 is a special case of inter-species PPI prediction in that there are significantly more known HIV-1 – human inter-species PPI than for any other virus or pathogen. Being able to predict novel PPI in under-studied viruses could be of significant

public benefit when there are new virus outbreaks, such as the recent Zika virus crisis. Another virus-host PPI prediction method made to predict PPI in under-studied viruses, Denovo [60], showed that the highest accuracy occurred when using inter-species interactions from similar viruses. PIPE was applied to Zika-human in [75], using PPI data between human and other viruses of the family (*Flaviviridae*) partly due to this finding. It would be useful to show that this finding does hold for PIPE and other host-pathogen PPI prediction methods.

For the case of cross-species PPI prediction, combining training data from multiple species should be examined further with PIPE. This research made changes that improved the accuracy of PIPE when using multiple training species. However, this research did not rigorously examine when different training species should be included in the training data at all; nor did it rigorously examine way of combining different species training data, for example by weighting the PIPE score of species differently based on evolutionary distance or number or quality of known interactions. Additionally, the different training species could be combined in a voting scheme, where multiple PIPE predictors, each trained on a different training species, are applied to the query protein pair. This research also showed that while closer species generally performed more accurately, this was not always the case. Part of the reason for this may be impacted by the reason that certain species are studied in the first place. For example, yeast species are often used as a way of studying DNA repair mechanisms, so many known PPI are related to these processes. As such, training with yeast-yeast interactions may perform more accurately on PPI in species with many known DNA-related PPI, even more so than closely related species. This could potentially be used to create training data sets that perform better for specific biological

processes, perhaps by combining PPI related to specific biological processes from a variety of different species.

The results of making predictions between SCN and soybean proteins were not analyzed for accuracy as there were no known interactions from which to test. However, these results could be analyzed by comparing to other PPI prediction methods, or by examining the biological pathways of the proteins in the top scoring pairs to see if proteins relating to the expected mechanisms of infection are predicted. The cross-species findings in this work could also be applied to other sequence-based methods to see if they hold. Meta methods created from multiple sequence-based prediction methods using many different species data could surpass any single algorithm for PPI prediction. Lastly, this research lays some ground work for transferring PPI prediction between different species. As more research is done and more PPI become known in a variety of different species, the potential of transferring the known PPI amongst the different species increases even further. The findings in this work could be re-examined once more interactions are known.

### 6.3.3 General PIPE Score Improvement

The normalization factor changes in Section 4.2.2 made a significant difference in the performance of cross-species PPI prediction, especially when there were many species used for the training data. The general effect of the normalization change was to reduce the size of the normalization factor (see Figure 24). This may indicate that the SW score could be improved in general by providing an upper limit on the normalization factor. The SW score for a pair of windows can never be larger than 1, even if there are many known PPI that support the interaction, which may cause reduced accuracy for general PPI prediction. This idea as well as the other ideas mentioned here should be explored in the future in order

to continue to improve PIPE and keep it at the leading edge of sequence-based PPI prediction

# References


[1] "Uses of Soybeans," *North Carolina Soybeans*, 2014. [Online]. Available: http://ncsoy.org/media-resources/uses-of-soybeans/. [Accessed: 30-Nov-2017].

[2] "10 Countries With Largest Soybean Production," *World Atlas*, 2017. [Online]. Available: https://www.worldatlas.com/articles/world-leaders-in-soya-soybean-production-by-country.html. [Accessed: 30-Nov-2017].

[3] "2017 Canadian Agricultural Outlook," *Agriculture and Agri-Food Canada*, 2017. [Online]. Available: http://www.agr.gc.ca/eng/about-us/publications/economic-publications/2017-canadian-agricultural-outlook.

[4] "Soybean Cyst Nematode Survey of Huron County," *Crop Adv. F. Crop Reports Results*, pp. 1–4, 2013.

[5] T. Allen *et al.*, "Soybean Yield Loss Estimates Due to Diseases in the United States and Ontario, Canada, from 2010 to 2014," *Plant Manag. Netw.*, 2017.

[6] C. Bradley *et al.*, "Soybean Disease Loss Estimates From the United States and Ontario, Canada — 2015," 2017.

[7] B. Mimee *et al.*, "First Report of Soybean Cyst Nematode (Heterodera glycines Ichinohe) on Soybean in the Province of Quebec, Canada," *APS Journals*, vol. 98, no. 3, 2014.

[8] G. L. Tylka and C. C. Marett, "Distribution of the Soybean Cyst Nematode, Heterodera glycines, in the United States and Canada: 1954 to 2014," *Plant Heal. Prog.*, vol. 15, no. 2, pp. 14–16, 2014.

[9] S. Pitre *et al.*, "PIPE: a protein-protein interaction prediction engine based on the re-occurring short polypeptide sequences between known interacting protein pairs.,"

*BMC Bioinformatics*, vol. 7, no. 365, 2006.

[10] A. Schoenrock, F. Dehne, J. R. Green, A. Golshani, and S. Pitre, "MP-PIPE: a massively parallel protein-protein interaction prediction engine," *Proc. Int. Conf. Supercomput.*, pp. 327–337, 2011.

[11] A. Schoenrock, "Realizing the Potential of Protein-Protein Interaction Prediction for Studying Single and Evolutionarily Similar Organisms and Engineering Inhibitory Proteins with InSiPS: The In Silico Protein Synthesizer," Carleton University, 2016.

[12] M. P. H. Stumpf *et al.*, "Estimating the size of the human interactome.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 105, no. 19, pp. 6959–64, 2008.

[13] K. Pruitt, G. Brown, T. Tatusova, and D. Maglott, "The Reference Sequence ( RefSeq ) Database," *NCBI Handb.*, pp. 1–24, 2002.

[14] S. Durmuş, T. Çakır, A. Özgür, and R. Guthke, "A review on computational systems biology of pathogen-host interactions.," *Front. Microbiol.*, vol. 6, no. April, p. 235, 2015.

[15] M. Jessulat *et al.*, "Recent advances in protein–protein interaction prediction: experimental and computational methods," *Expert Opin. Drug Discov.*, vol. 6, no. 9, pp. 921–935, 2011.

[16] C. von Mering *et al.*, "Comparative assessment of large-scale data sets of protein–protein interactions," *Nature*, vol. 417, no. 6887, pp. 399–403, 2002.

[17] S. H. Park, J. M. Goo, and C.-H. Jo, "Receiver operating characteristic (ROC) curve: practical review for radiologists.," *Korean J. Radiol.*, vol. 5, no. March, pp. 11–8, 2004.

[18] O. Keskin, N. Tuncbag, and A. Gursoy, "Predicting Protein-Protein Interactions

from the Molecular to the Proteome Level," *Chem. Rev.*, vol. 116, no. 8, pp. 4884–4909, 2016.

[19] J. Zhang and L. Kurgan, "Review and comparative assessment of sequence-based predictors of protein-binding residues," *Brief. Bioinform.*, vol. bbx022, no. 2, pp. 1–17, 2017.

[20] A. Chatr-Aryamontri *et al.*, "The BioGRID interaction database: 2015 update," *Nucleic Acids Res.*, vol. 43, no. Database Issue, pp. D470–D478, 2015.

[21] A. Chatr-Aryamontri *et al.*, "MINT: the Molecular INTeraction database," *Nucleic Acids Res.*, vol. 35, no. Database Issue, pp. 572–574, 2007.

[22] S. Kerrien *et al.*, "The IntAct molecular interaction database in 2012," *Nucleic Acids Res.*, vol. 40, no. D1, pp. 841–846, 2012.

[23] I. Xenarios, "DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions," *Nucleic Acids Res.*, vol. 30, no. 1, pp. 303–305, 2002.

[24] T. S. Keshava Prasad *et al.*, "Human Protein Reference Database - 2009 update," *Nucleic Acids Res.*, vol. 37, no. SUPPL. 1, pp. 767–772, 2009.

[25] U. Guldener, "MPact: the MIPS protein interaction resource on yeast," *Nucleic Acids Res.*, vol. 34, no. 90001, pp. D436–D441, 2006.

[26] J. Wu, T. Vallenius, K. Ovaska, J. Westermarck, T. P. Mäkelä, and S. Hautaniemi, "Integrated network analysis platform for protein-protein interactions," *Nat. Methods*, vol. 6, no. 1, pp. 75–77, 2009.

[27] A. Calderone, L. Licata, and G. Cesareni, "VirusMentha: a new resource for virus-host protein interactions.," *Nucleic Acids Res.*, vol. 43, no. D1, pp. 1–5, 2014.

[28] D. Ako-Adjei *et al.*, "HIV-1, human interaction database: current status and new features.," *Nucleic Acids Res.*, vol. 43, no. Database issue, pp. D566-70, 2015.

[29] Y. Park and E. M. Marcotte, "Flaws in evaluation schemes for pair-input computational predictions," *Nat. meth*, vol. 9, no. 12, pp. 1134–1136, 2012.

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

[31] P. Blohm *et al.*, "Negatome 2.0: A database of non-interacting proteins derived by literature mining, manual annotation and protein structure analysis," *Nucleic Acids Res.*, vol. 42, no. D1, pp. 396–400, 2014.

[32] J. Zahiri, J. H. Bozorgmehr, and A. Masoudi-Nejad, "Computational Prediction of Protein–Protein Interaction Networks: Algorithms and Resources," *Curr. Genomics*, vol. 14, no. 6, pp. 397–414, 2013.

[33] E. L. Folador *et al.*, "An improved interolog mapping-based computational prediction of protein–protein interactions with increased network coverage," *Integr. Biol.*, vol. 6, no. 11, pp. 1080–1087, 2014.

[34] D. S. Goldberg and F. P. Roth, "Assessing experimentally derived interactions in a small world," *Proc. Natl. Acad. Sci.*, vol. 100, no. 8, pp. 4372–4376, 2003.

[35] S. Wuchty, "Topology and weights in a protein domain interaction network--a novel way to predict protein interactions.," *BMC Genomics*, vol. 7, p. 122, 2006.

[36] T. Dandekar, B. Snel, M. Huynen, and P. Bork, "Conservation of gene order: A fingerprint of proteins that physically interact," *Trends Biochem. Sci.*, vol. 23, no. 9, pp. 324–328, 1998.

[37] A. J. Enright, I. Illopoulos, N. C. Kyrpides, and C. A. Ouzounis, "Protein interaction

maps for complete genomes based on gene fusion events," *Nature*, vol. 402, no. 6757, pp. 86–90, 1999.

[38] Y. Li and L. Ilie, "SPRINT: Ultrafast protein-protein interaction prediction of the entire human interactome," *BMC Bioinformatics*, vol. 18, no. 1, pp. 1–11, 2017.

[39] S. Martin, D. Roe, and J. L. Faulon, "Predicting protein-protein interactions using signature products," *Bioinformatics*, vol. 21, no. 2, pp. 218–226, 2005.

[40] 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. 4337–41, 2007.

[41] Y. Guo, L. Yu, Z. Wen, and M. Li, "Using support vector machine combined with auto covariance to predict protein-protein interactions from protein sequences," *Nucleic Acids Res.*, vol. 36, no. 9, pp. 3025–3030, 2008.

[42] Y. Park, "Critical assessment of sequence-based protein-protein interaction prediction methods that do not require homologous protein sequences," *BMC Bioinformatics*, vol. 10, no. 1, p. 419, 2009.

[43] Y. Ding, J. Tang, and F. Guo, "Predicting protein-protein interactions via multivariate mutual information of protein sequences," *BMC Bioinformatics*, vol. 17, no. 1, p. 398, 2016.

[44] M. Li, B. Ma, D. Kisman, and J. Tromp, "PatternHunter II: highly sensitive and fast homology search.," *Genome Inform.*, vol. 14, no. 3, pp. 164–75, 2003.

[45] R. Schwartz and M. Dayhoff, "Matrices for detecting distant relationships," in *Atlas of Protein Sequence and Structure*, Supplement., Silver Spring, MD: National Biomedical Research Foundation, 1978, pp. 353–358.

[46] L. Ilie, S. Ilie, and A. M. Bigvand, "SpEED: Fast computation of sensitive spaced

seeds," *Bioinformatics*, vol. 27, no. 17, pp. 2433–2434, 2011.

[47] E. Nourani, F. Khunjush, and S. Durmuş, "Computational approaches for prediction of pathogen-host protein-protein interactions.," *Front. Microbiol.*, vol. 6, no. 94, pp. 1–10, 2015.

[48] O. Tastan, Y. Qi, J. G. Carbonell, and J. Klein-Seetharaman, "Prediction of Interactions Between Hiv-1 and Human Proteins By Information Integration," in *Pacific Symposium on Biocomputing*, 2009, pp. 516–527.

[49] Y. Qi, O. Tastan, J. G. Carbonell, J. Klein-Seetharaman, and J. Weston, "Semi-supervised multi-task learning for predicting interactions between HIV-1 and human proteins," *Bioinformatics*, vol. 27, no. 13, pp. i645–i652, 2011.

[50] P. Evans, W. Dampier, L. Ungar, and A. Tozeren, "Prediction of HIV-1 virus-host protein interactions using virus and host sequence motifs.," *BMC Med. Genomics*, vol. 2, p. 27, 2009.

[51] M. D. Dyer, T. M. Murali, and B. W. Sobral, "Supervised learning and prediction of physical interactions between human and HIV proteins," *Infect. Genet. Evol.*, vol. 11, no. 5, pp. 917–923, 2011.

[52] S. Mei, "Probability weighted ensemble transfer learning for predicting interactions between HIV-1 and human proteins," *PLoS One*, vol. 8, no. 11, pp. 1–13, 2013.

[53] I. Nouretdinov, A. Gammerman, Y. Qi, and J. Klein-Seetharaman, "Determining confidence of predicted interactions between HIV-1 and human proteins using conformal method.," in *Pacific Symposium on Biocomputing*, 2012, pp. 311–22.

[54] A. Mukhopadhyay, S. Ray, and U. Maulik, "Incorporating the type and direction information in predicting novel regulatory interactions between HIV-1 and human

proteins using a biclustering approach.," *BMC Bioinformatics*, vol. 15, p. 26, 2014.

[55] J. M. Doolittle and S. M. Gomez, "Structural similarity-based predictions of protein interactions between HIV-1 and Homo sapiens.," *Virol. J.*, vol. 7, p. 82, 2010.

[56] C. Zhao and A. Sacan, "Prediction of HIV-1 and human protein interactions based on a novel evolution-aware structure alignment method," in *Proceedings of the International Conference on Bioinformatics & Computational Biology (BIOCOMP)*, 2013, pp. 1–7.

[57] A. Emamjomeh, B. Goliaei, J. Zahiri, and R. Ebrahimpour, "Predicting protein-protein interactions between human and hepatitis C virus via an ensemble learning method.," *Mol. Biosyst.*, vol. 10, no. 12, pp. 3147–54, 2014.

[58] D. Mairiang *et al.*, "Identification of New Protein Interactions between Dengue Fever Virus and Its Hosts, Human and Mosquito," *PLoS One*, vol. 8, no. 1, 2013.

[59] S. S. Sahu, T. Weirick, and R. Kaundal, "Predicting genome-scale Arabidopsis-Pseudomonas syringae interactome using domain and interolog-based approaches," *BMC Bioinformatics*, vol. 15, no. Suppl 11, p. S13, 2014.

[60] F.-E. Eid, M. ElHefnawi, and L. S. Heath, "DeNovo: Virus-Host Sequence-Based Protein-Protein Interaction Prediction," *Bioinforma.* , pp. 1–7, 2015.

[61] S. Yang, H. Li, H. He, Y. Zhou, and Z. Zhang, "Critical assessment and performance improvement of plant–pathogen protein–protein interaction prediction methods," *Brief. Bioinform.*, no. August, pp. 1–14, 2017.

[62] X. W. Chen, M. Liu, and R. Ward, "Protein function assignment through mining cross-species protein-protein interactions," *PLoS One*, vol. 3, no. 2, 2008.

[63] S. Pitre *et al.*, "Global investigation of protein-protein interactions in yeast

Saccharomyces cerevisiae using re-occurring short polypeptide sequences," *Nucleic Acids Res.*, vol. 36, no. 13, pp. 4286–4294, 2008.

[64] C. Patulea, "Targeted Optimization of Computational and Classification Performance of a Protein-Protein Interaction Predictor," Carleton University, 2011.

[65] S. Pitre *et al.*, "Short Co-occurring Polypeptide Regions Can Predict Global Protein Interaction Maps," *Sci. Rep.*, vol. 2, pp. 1–10, 2012.

[66] A. Amos-Binks *et al.*, "Binding site prediction for protein-protein interactions and novel motif discovery using re-occurring polypeptide sequences.," *BMC Bioinformatics*, vol. 12, no. 1, p. 225, 2011.

[67] A. Schoenrock *et al.*, "Efficient prediction of human protein-protein interactions at a global scale," *BMC Bioinformatics*, vol. 15, no. 1, pp. 1–22, 2014.

[68] T. U. Consortium, "UniProt: a hub for protein information," *Nucleic Acids Res.*, vol. 43, no. D1, pp. D204–D212, 2015.

[69] S. Kumar and S. B. Hedges, "Timetree2: Species divergence times on the iPhone," *Bioinformatics*, vol. 27, no. 14, pp. 2023–2024, 2011.

[70] M. Hollander, D. A. Wolfe, and E. Chicken, *Nonparametric Statistical Methods*, 3rd ed. New York: John Wiley & Sons, 2013.

[71] B. Barnes *et al.*, "Predicting Novel Protein-Protein Interactions Between the HIV-1 Virus and Homo Sapiens," in *Student Conference (ISC), 2016 IEEE EMBS International*, 2016, pp. 1–4.

[72] K. D. Pruitt *et al.*, "RefSeq: An update on mammalian reference sequences," *Nucleic Acids Res.*, vol. 42, no. D1, pp. 756–763, 2014.

[73] T. Kazmirchuk *et al.*, "Designing anti-Zika virus peptides derived from predicted

human-Zika virus protein-protein interactions," *Comput. Biol. Chem.*, vol. 71, pp. 180–187, 2017.

[74] A. Schoenrock, D. Burnside, A. Wong, and F. Dehne, "Engineering inhibitory proteins with InSiPS: the in-silico protein synthesizer," in *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, 2015, pp. 25–35.

[75] T. Kazmirchuk *et al.*, "Designing anti-Zika virus peptides derived from predicted human-Zika virus protein-protein interactions," *Comput. Biol. Chem.*, vol. 71, pp. 180–187, 2017.

# Appendix A: Cross-species experiments

Appendix A includes plots and tables that were too numerous to include in Chapter 4: Cross-species PPI prediction. Here, we present all cross-species PPI prediction curves for the normalization factor change, and the scoring metrics for each training species test species pair. We then include the P-R curves for each test species in the evolutionary distance association relation to accuracy section, as well as the performance ranks for each scoring metrics.

## A.1 Cross species score normalization changes

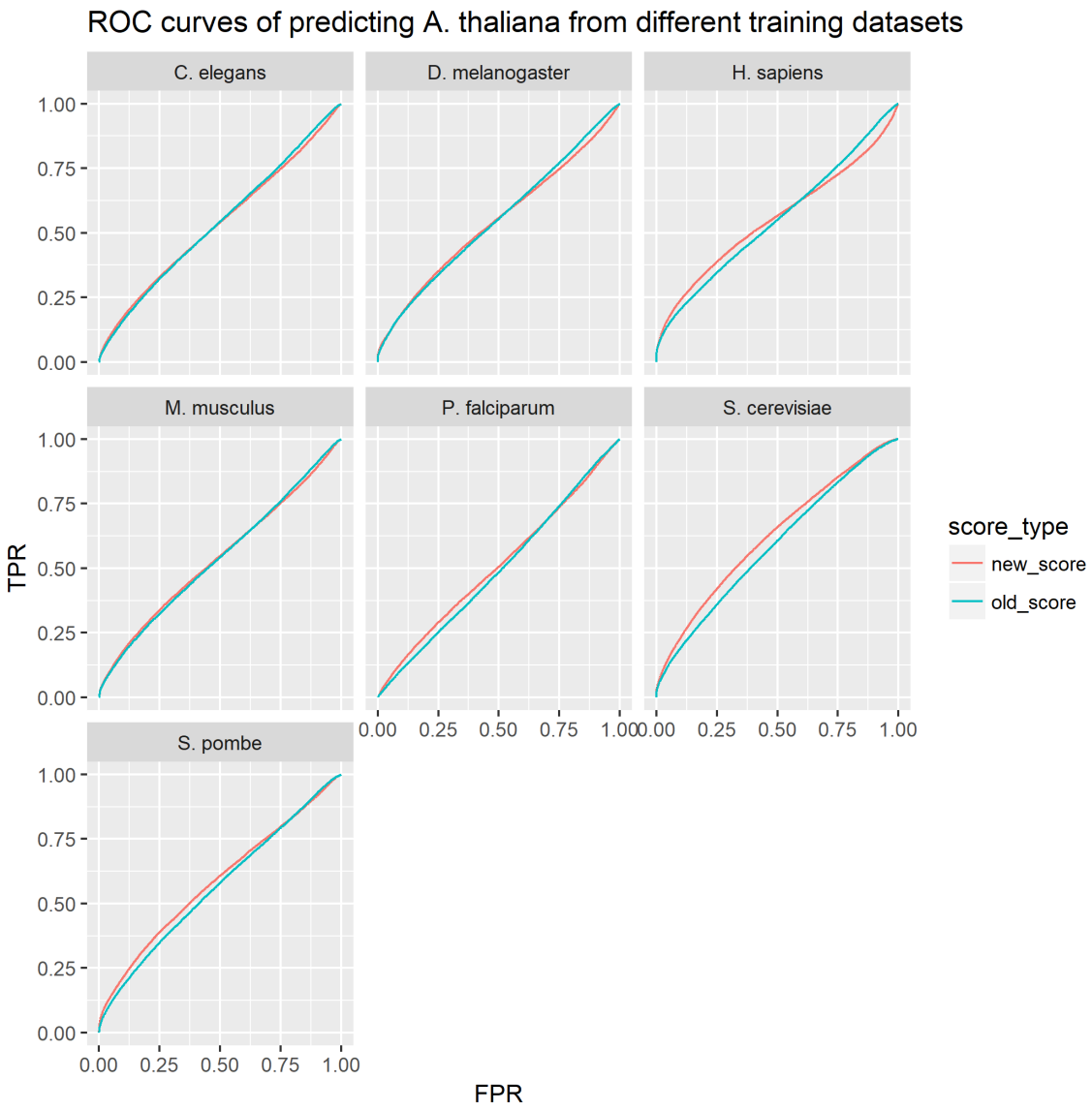



**Figure 37: ROC curve for predicting A. thaliana from every other training species. Subplot title gives the training species.**

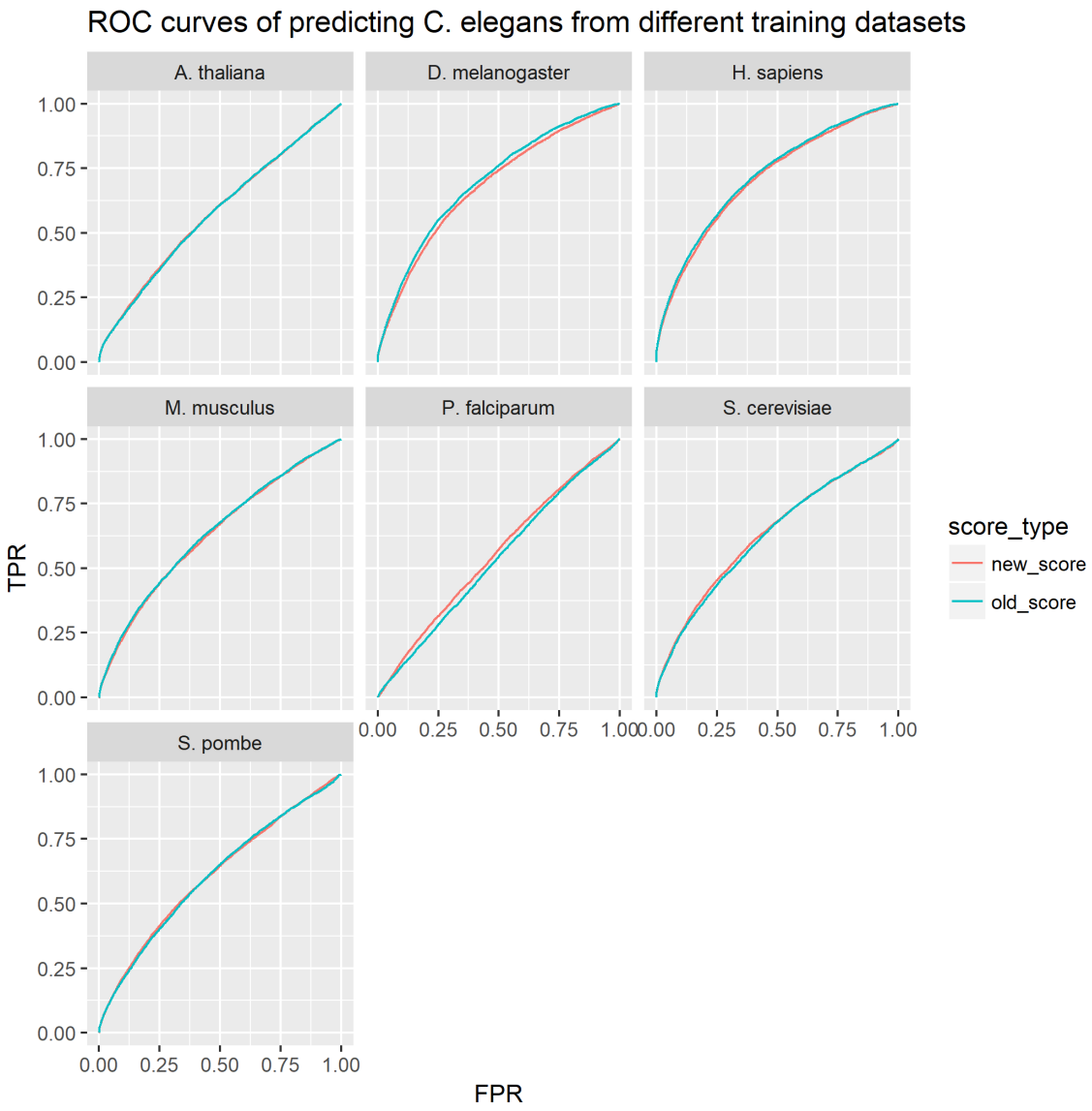



**Figure 38: ROC curve for predicting C. elegans from every other training species. Subplot title gives the training species.**

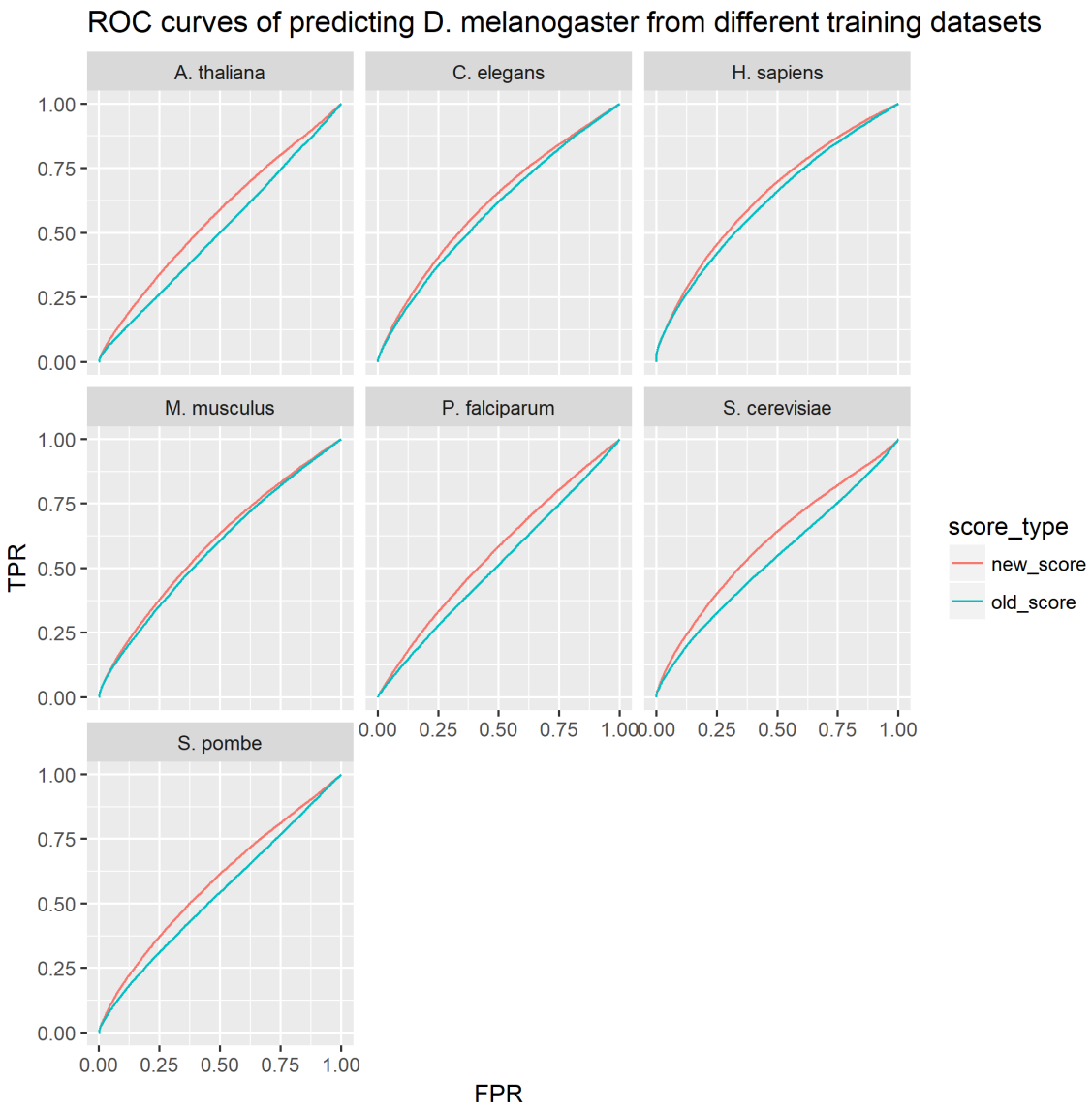



**Figure 39: ROC curve for predicting D. melanogaster from every other training species. Subplot title gives the training species.**

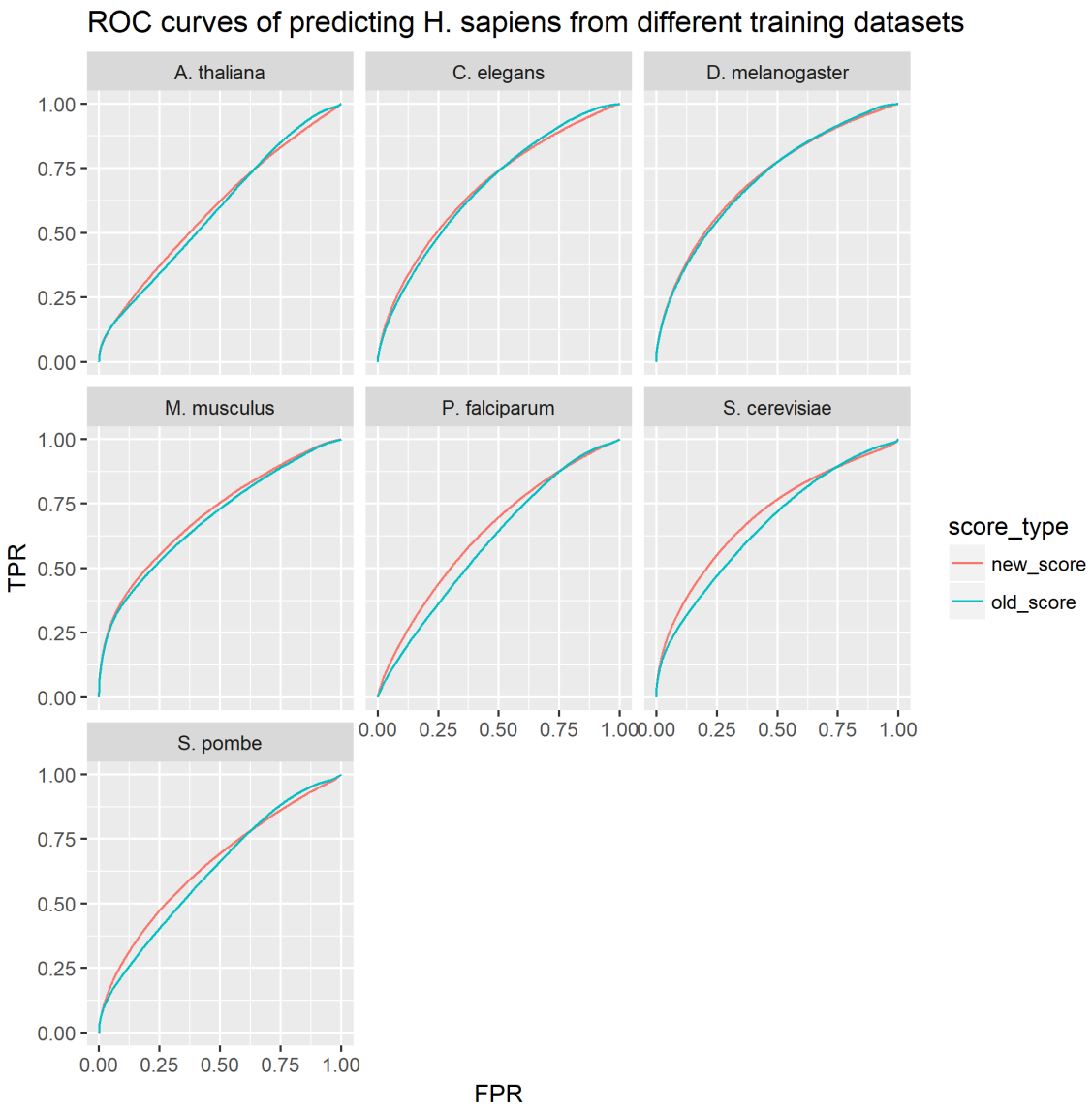



**Figure 40: ROC curve for predicting H. sapiens from every other training species. Subplot title gives the training species.**

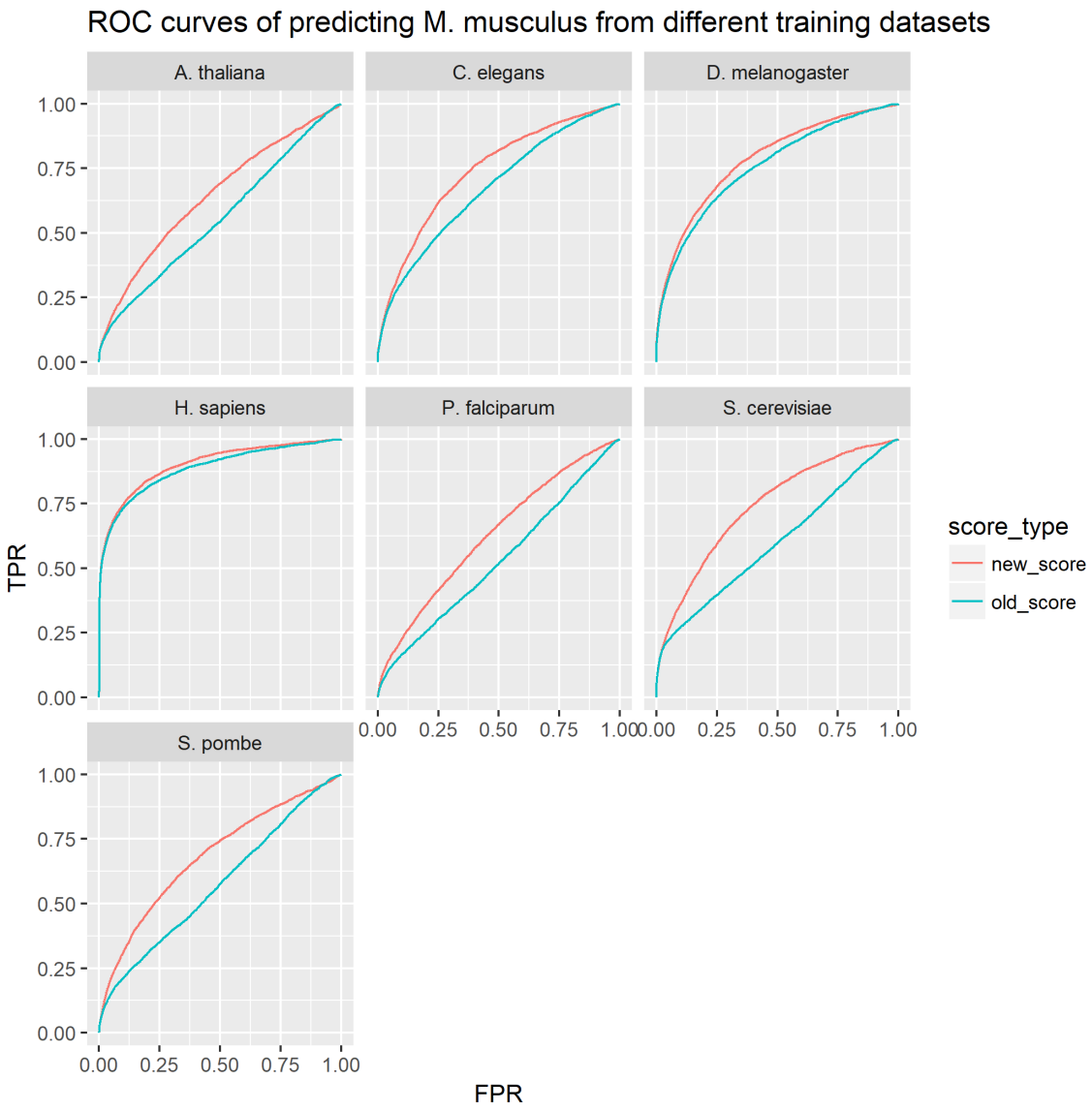



**Figure 41: ROC curve for predicting M. musculus from every other training species. Subplot title gives the training species.**

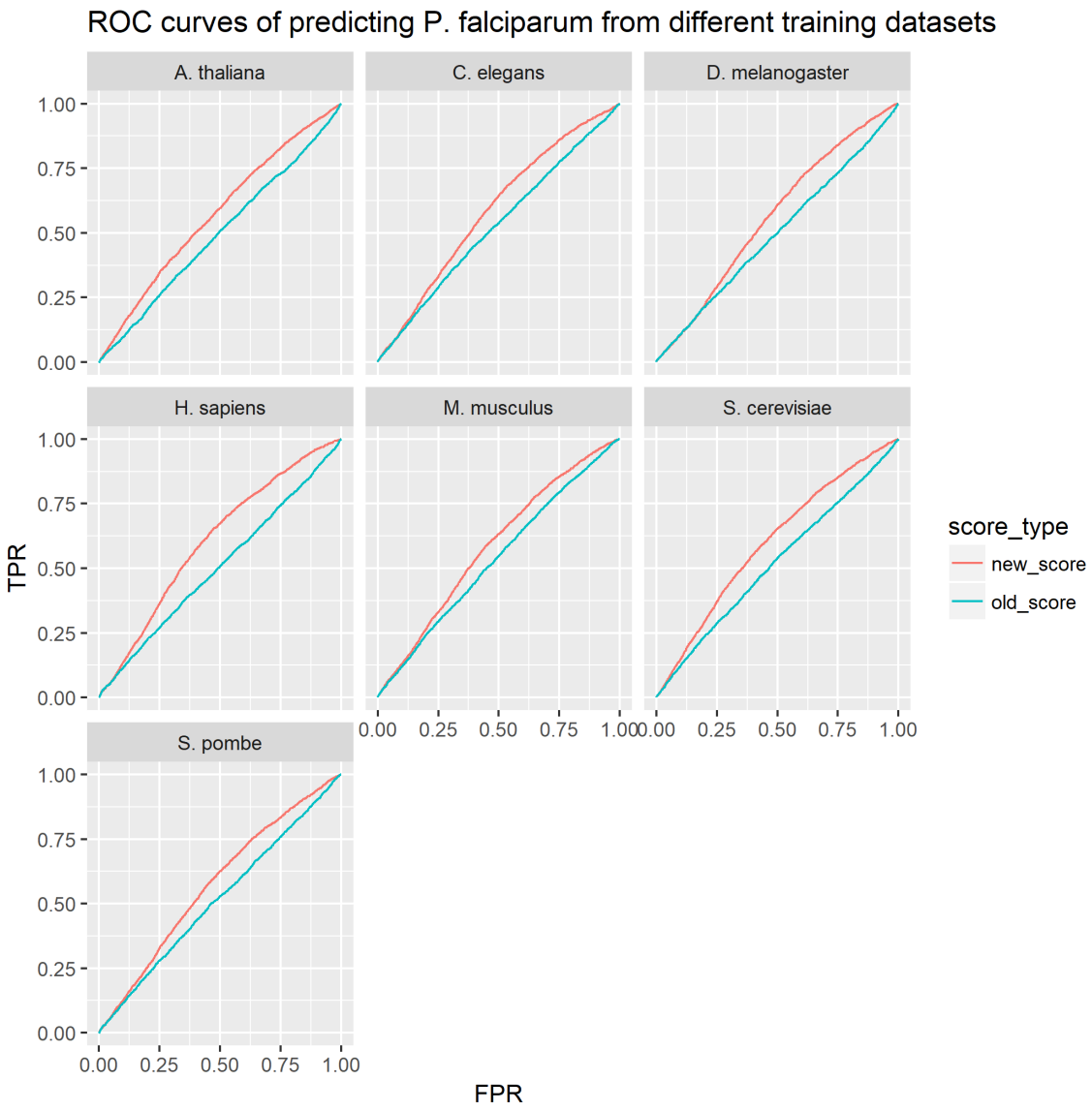



**Figure 42: ROC curve for predicting P. falciparum from every other training species. Subplot title gives the training species.**

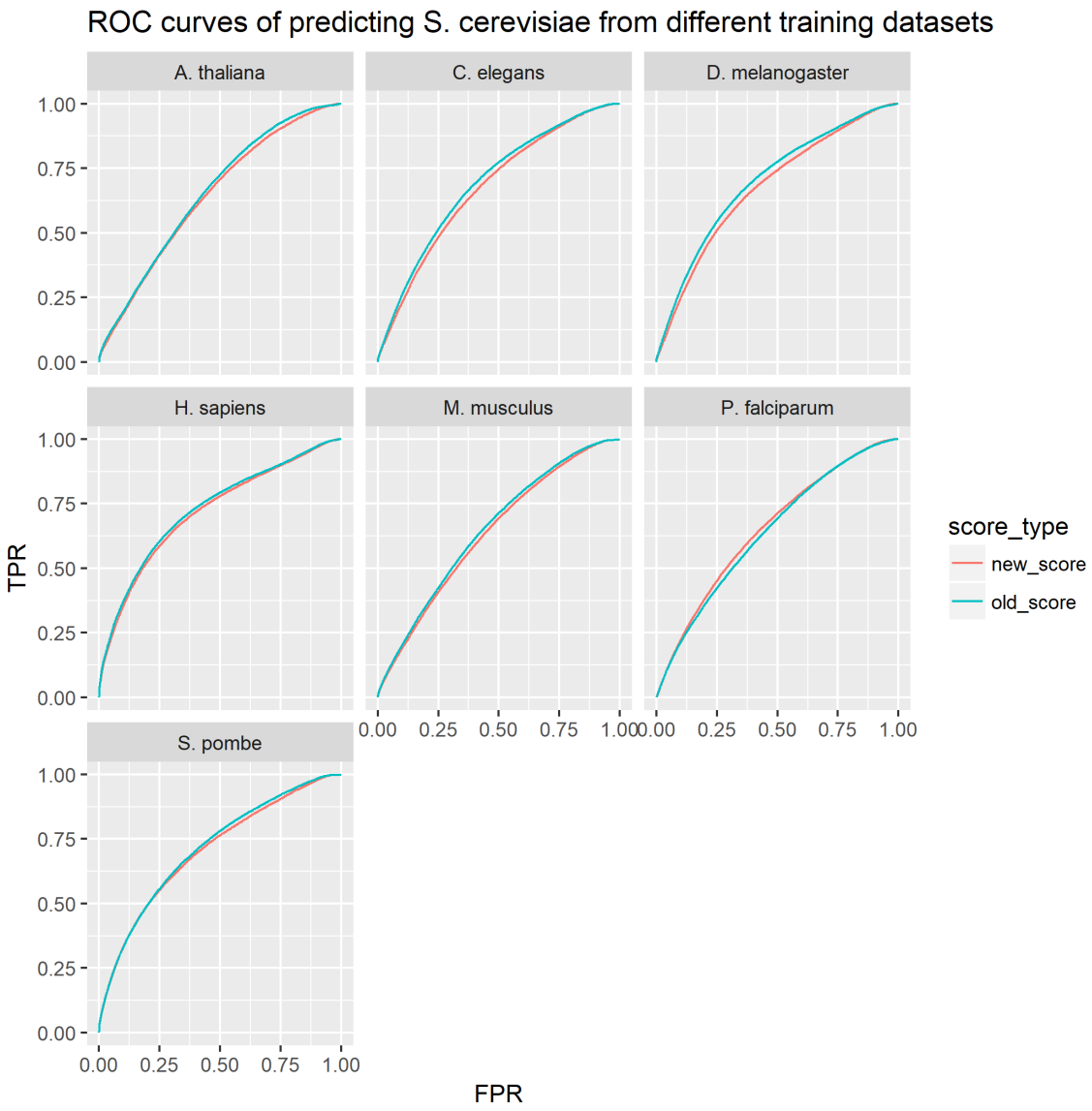



**Figure 43: ROC curve for predicting S. cerevisiae from every other training species. Subplot title gives the training species.**

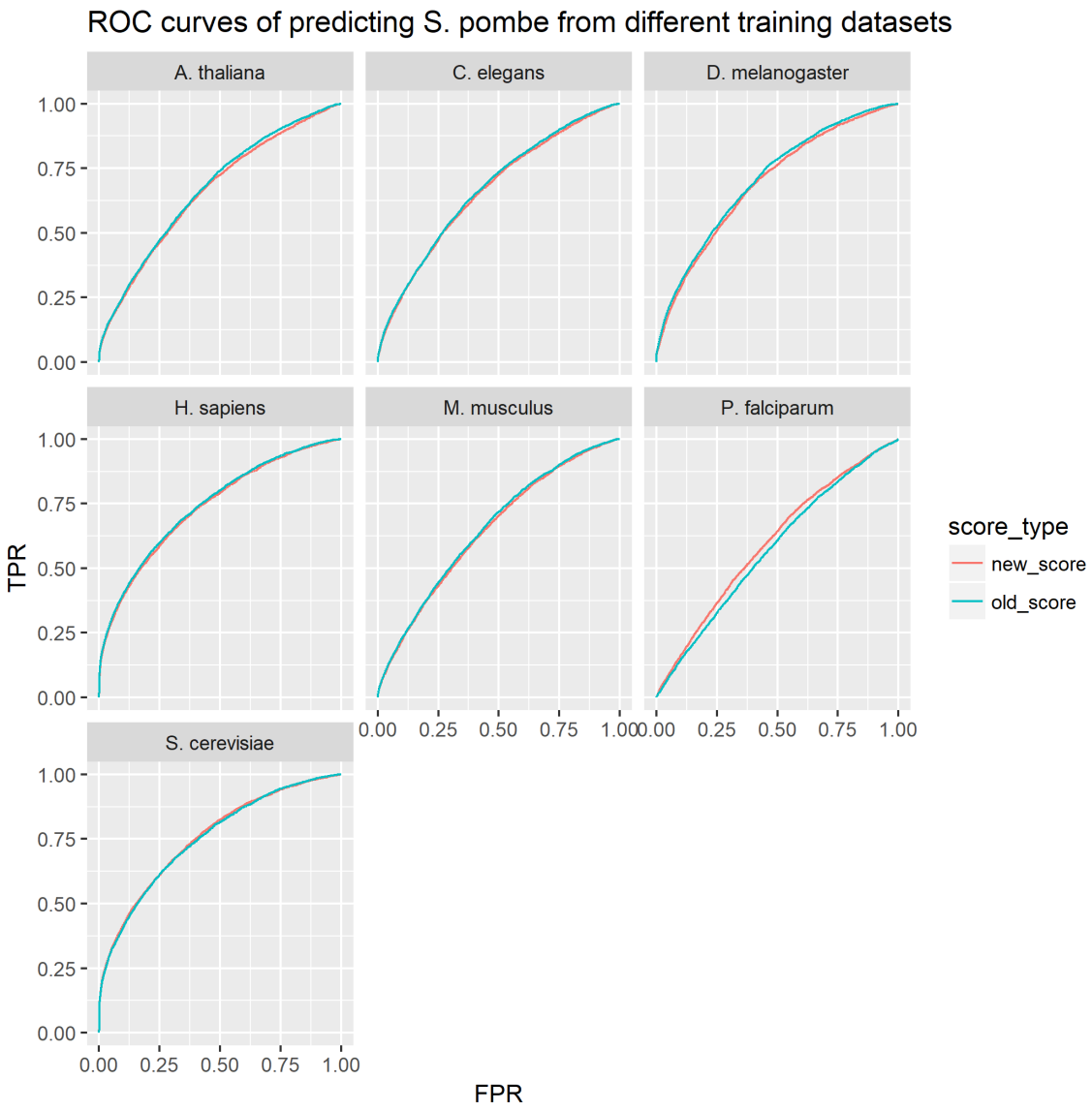



**Figure 44: ROC curve for predicting S. pombe from every other training species. Subplot title gives the training species.**

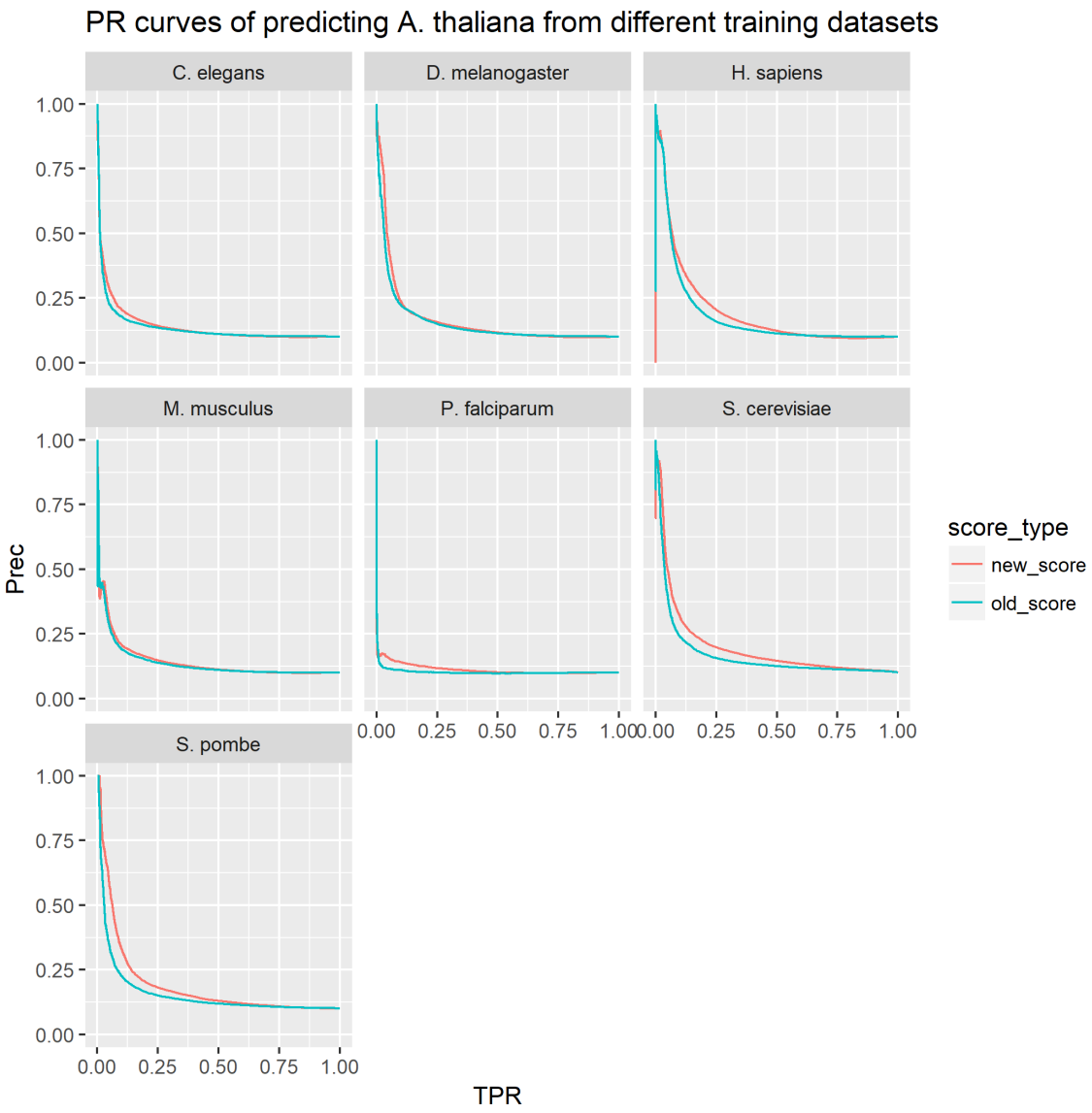



**Figure 45: P-R curve for predicting A. thaliana from every other training species (at 10:1 CI). Subplot title gives the training species.**

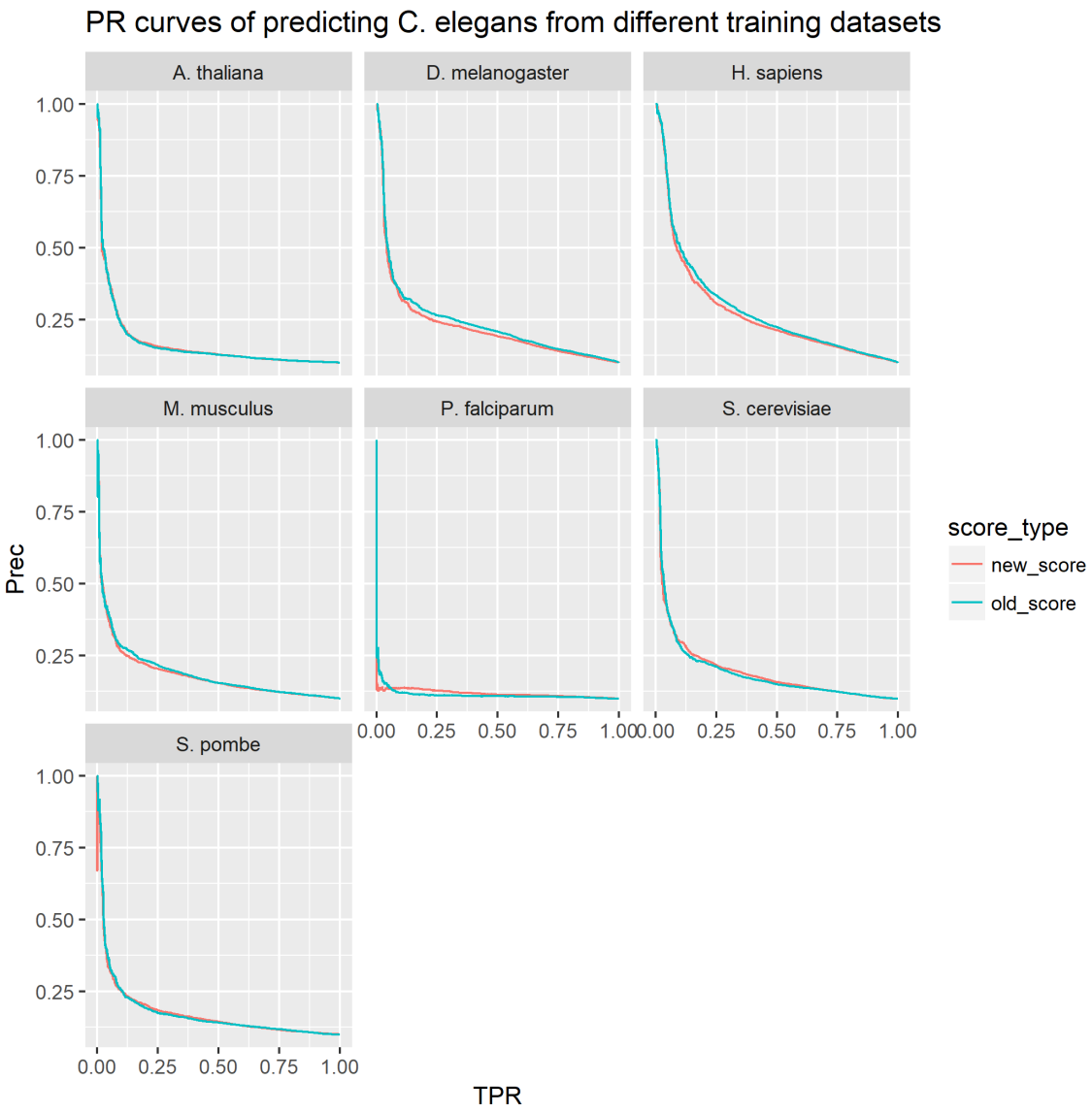



**Figure 46: P-R curve for predicting C. elegans from every other training species (at 10:1 CI). Subplot title gives the training species.**

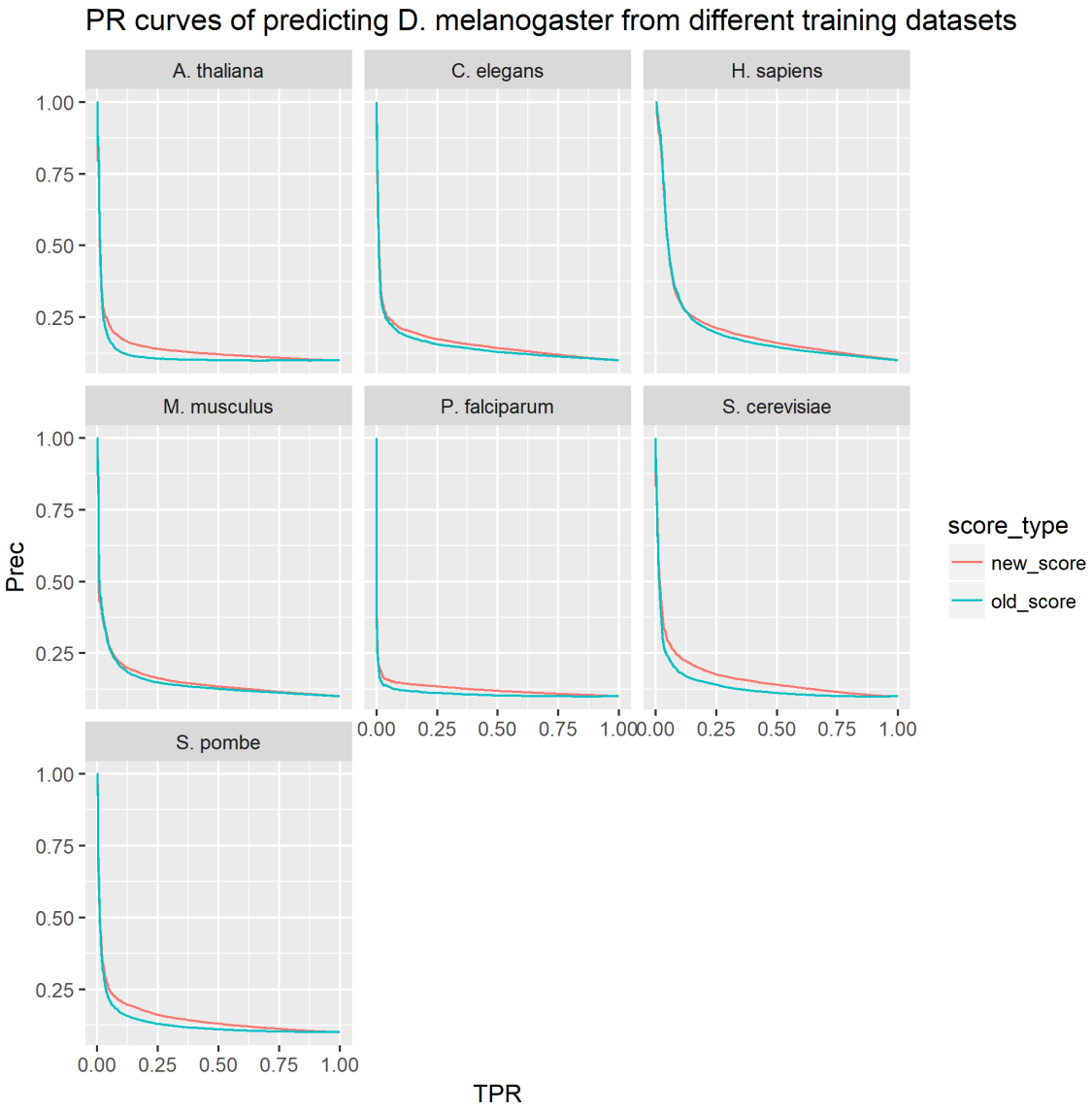



**Figure 47: P-R curve for predicting *D. melanogaster* from every other training species (at 10:1 CI). Subplot title gives the training species.**

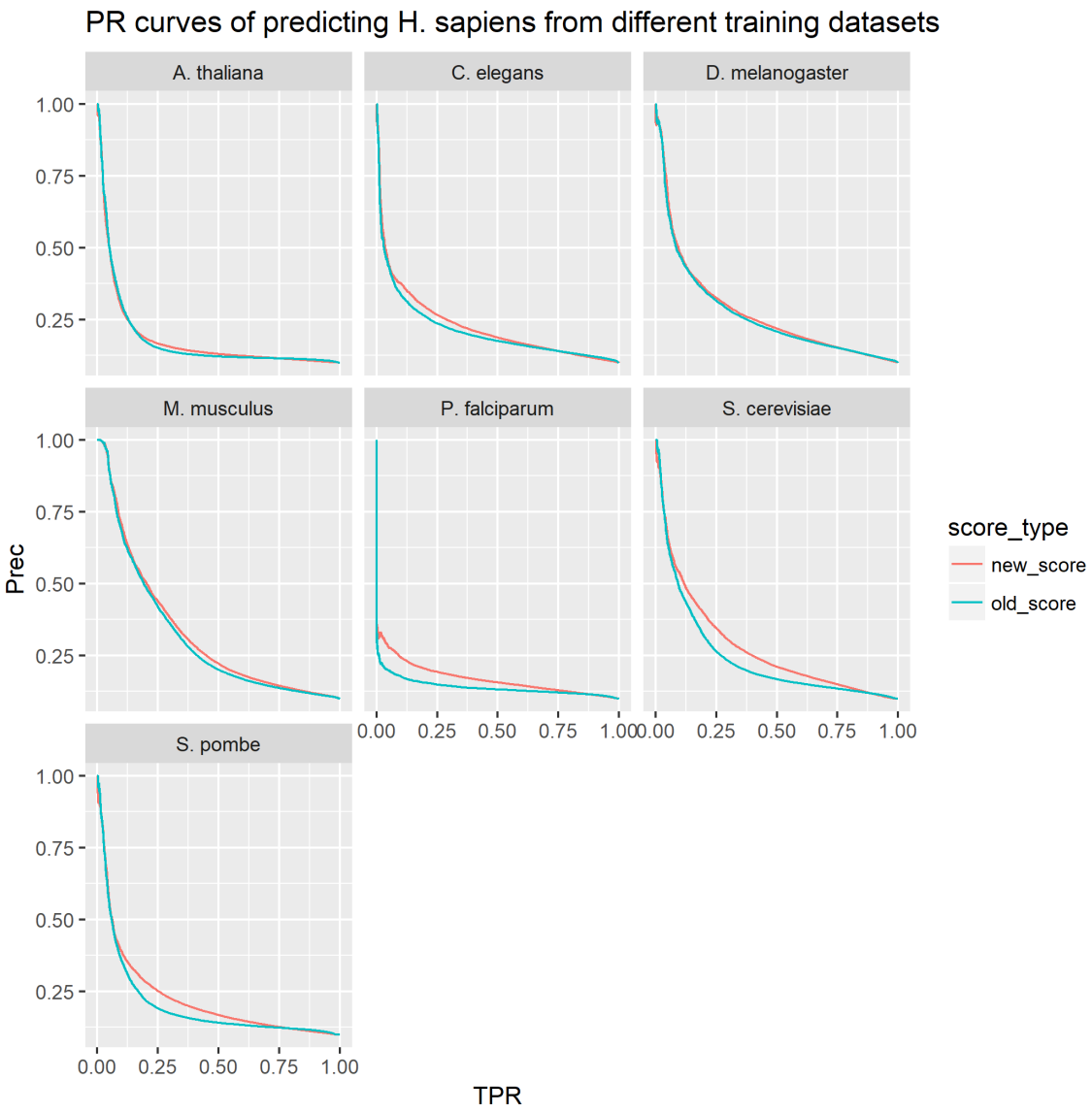



**Figure 48: P-R curve for predicting *H. sapiens* from every other training species (at 10:1 CI). Subplot title gives the training species.**

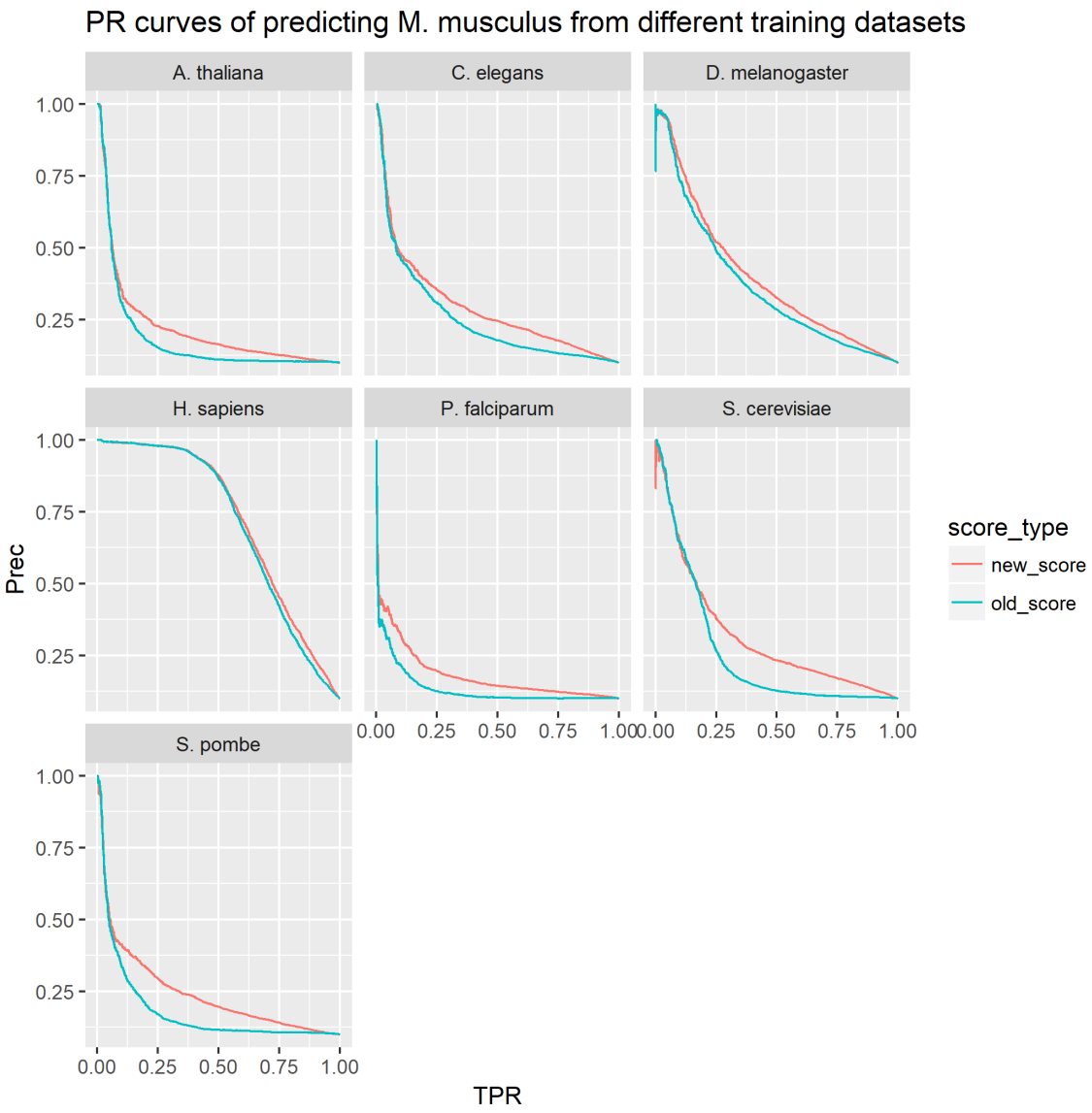



**Figure 49: P-R curve for predicting *M. musculus* from every other training species (at 10:1 CI). Subplot title gives the training species.**

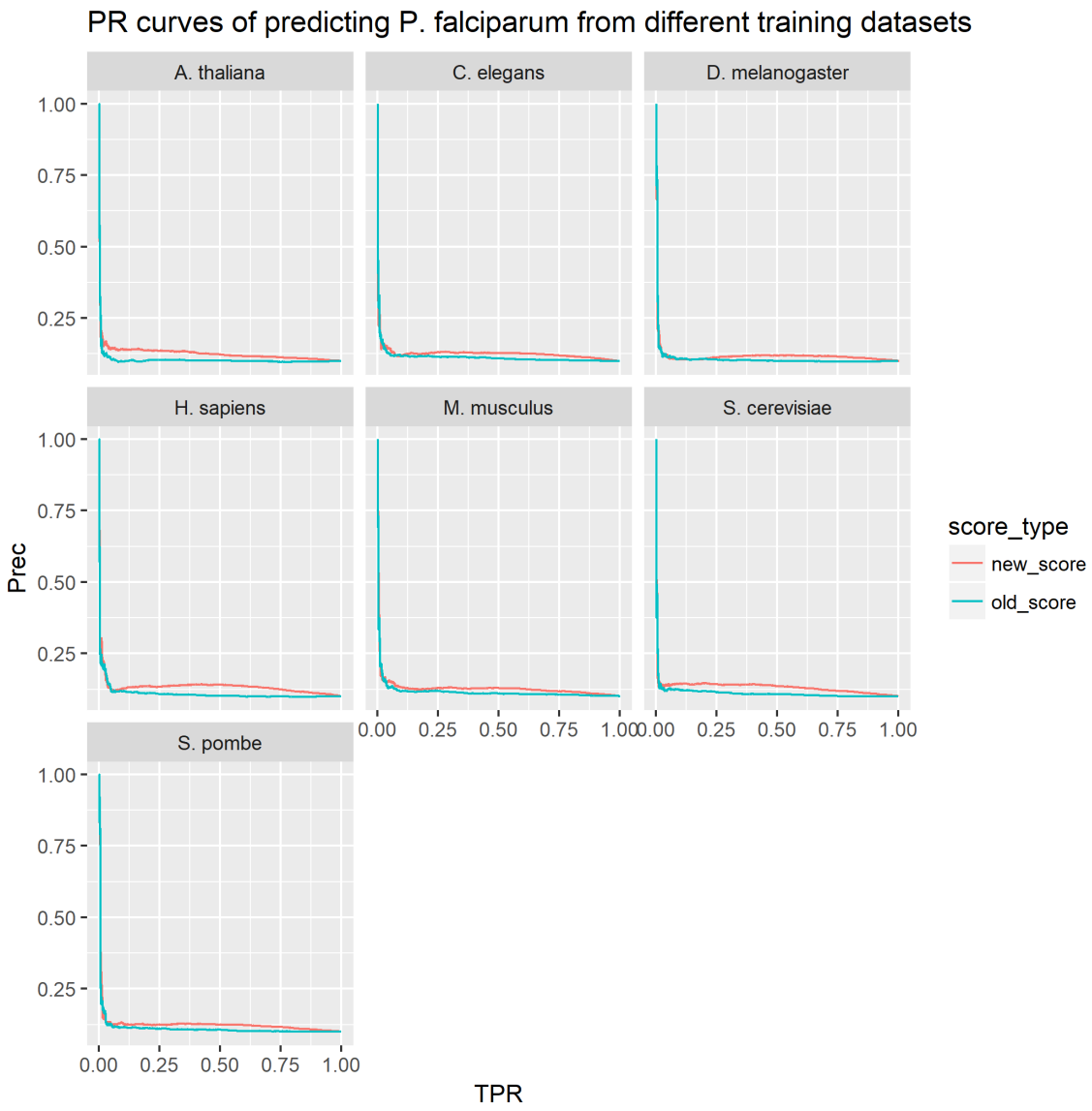



**Figure 50: P-R curve for predicting *P. falciparum* from every other training species (at 10:1 CI). Subplot title gives the training species.**

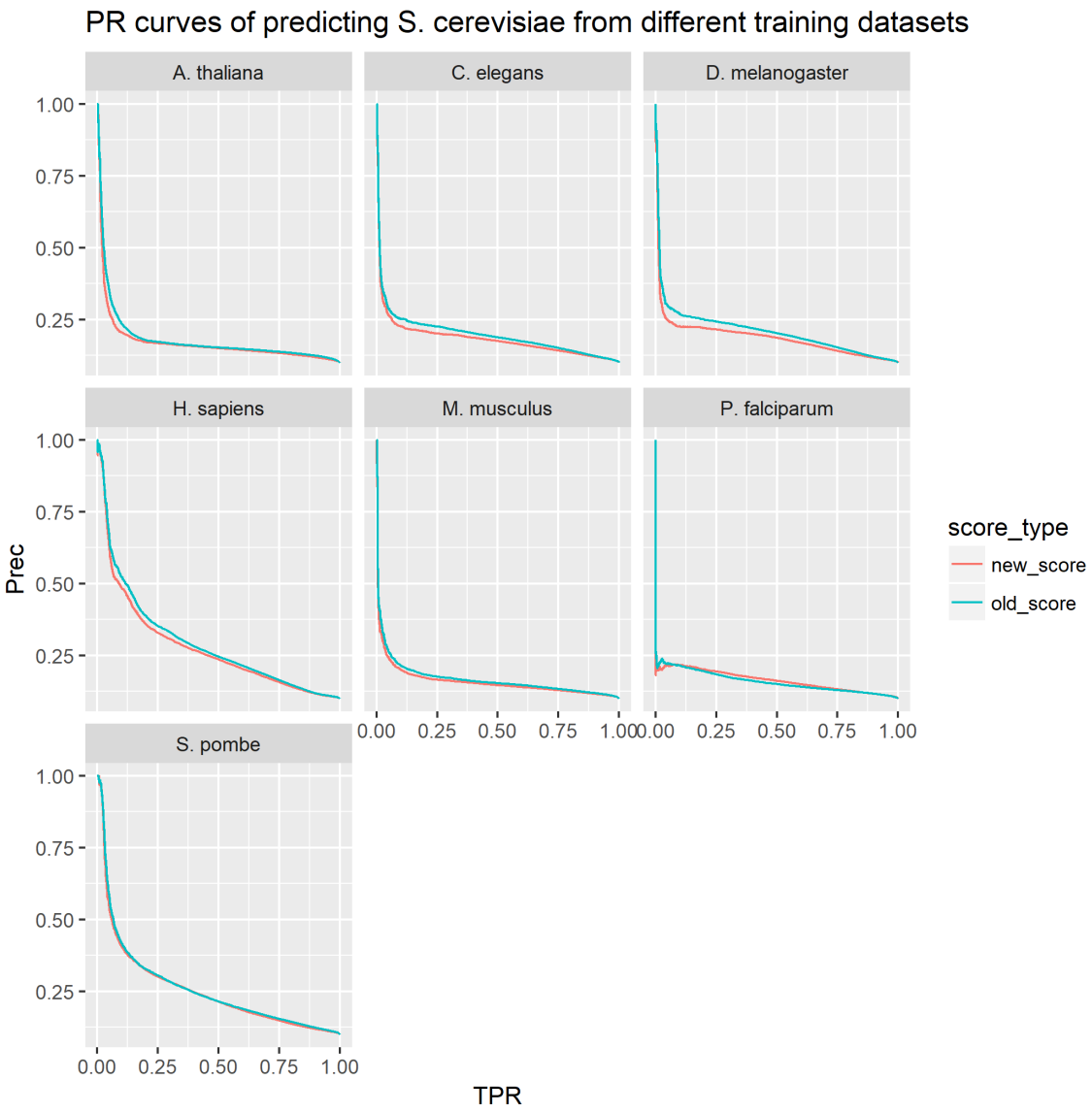



**Figure 51: P-R curve for predicting *S. cerevisiae* from every other training species (at 10:1 CI). Subplot title gives the training species.**

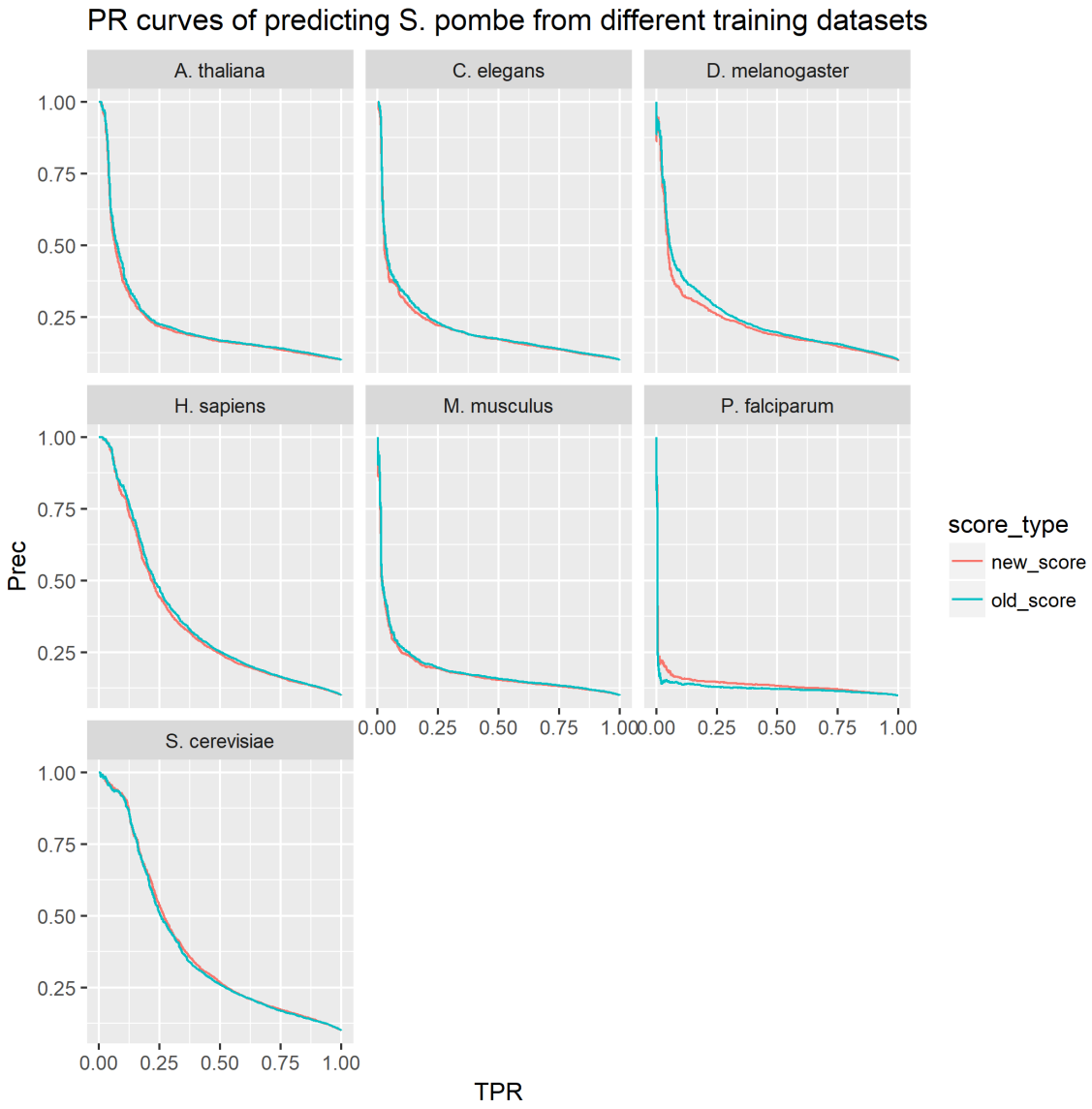



**Figure 52: P-R curve for predicting *S. pombe* from every other training species (at 10:1 CI). Subplot title gives the training species.**

**Table 18: Single-species cross-species prediction results for two performance metrics (evaluated at 10:1 CI) for every pair of training and testing species.**

| Testing species | Training species | AU-PRC | | Prec at 25% TPR | |
|---|---|---|---|---|---|
| | | New | Old | New | Old |
| *A. thaliana* | *C. elegans* | 0.141 | 0.135 | 0.142 | 0.135 |
| *A. thaliana* | *D. melanogaster* | 0.165 | 0.155 | 0.157 | 0.150 |
| *A. thaliana* | *H. sapiens* | 0.196 | 0.180 | 0.205 | 0.160 |
| *A. thaliana* | *M. musculus* | 0.144 | 0.139 | 0.147 | 0.138 |
| *A. thaliana* | *P. falciparum* | 0.113 | 0.103 | 0.118 | 0.101 |
| *A. thaliana* | *S. cerevisiae* | 0.197 | 0.168 | 0.199 | 0.157 |

| | | | | | |
|---|---|---|---|---|---|
| *A. thaliana* | *S. pombe* | 0.189 | 0.158 | 0.182 | 0.151 |
| *C. elegans* | *A. thaliana* | 0.162 | 0.161 | 0.156 | 0.151 |
| *C. elegans* | *D. melanogaster* | 0.223 | 0.237 | 0.243 | 0.265 |
| *C. elegans* | *H. sapiens* | 0.270 | 0.283 | 0.307 | 0.335 |
| *C. elegans* | *M. musculus* | 0.183 | 0.188 | 0.204 | 0.217 |
| *C. elegans* | *P. falciparum* | 0.118 | 0.113 | 0.128 | 0.111 |
| *C. elegans* | *S. cerevisiae* | 0.196 | 0.192 | 0.218 | 0.212 |
| *C. elegans* | *S. pombe* | 0.175 | 0.174 | 0.185 | 0.175 |
| *D. melanogaster* | *A. thaliana* | 0.139 | 0.118 | 0.140 | 0.106 |
| *D. melanogaster* | *C. elegans* | 0.159 | 0.148 | 0.174 | 0.156 |
| *D. melanogaster* | *H. sapiens* | 0.206 | 0.197 | 0.213 | 0.196 |
| *D. melanogaster* | *M. musculus* | 0.155 | 0.148 | 0.164 | 0.148 |
| *D. melanogaster* | *P. falciparum* | 0.124 | 0.109 | 0.135 | 0.112 |
| *D. melanogaster* | *S. cerevisiae* | 0.165 | 0.138 | 0.177 | 0.141 |
| *D. melanogaster* | *S. pombe* | 0.151 | 0.131 | 0.160 | 0.129 |
| *H. sapiens* | *A. thaliana* | 0.180 | 0.177 | 0.166 | 0.151 |
| *H. sapiens* | *C. elegans* | 0.225 | 0.212 | 0.266 | 0.237 |
| *H. sapiens* | *D. melanogaster* | 0.273 | 0.266 | 0.326 | 0.316 |
| *H. sapiens* | *M. musculus* | 0.324 | 0.311 | 0.438 | 0.419 |
| *H. sapiens* | *P. falciparum* | 0.169 | 0.140 | 0.193 | 0.149 |
| *H. sapiens* | *S. cerevisiae* | 0.277 | 0.242 | 0.345 | 0.265 |
| *H. sapiens* | *S. pombe* | 0.223 | 0.200 | 0.251 | 0.191 |
| *M. musculus* | *A. thaliana* | 0.217 | 0.179 | 0.227 | 0.154 |
| *M. musculus* | *C. elegans* | 0.291 | 0.249 | 0.353 | 0.310 |
| *M. musculus* | *D. melanogaster* | 0.395 | 0.365 | 0.520 | 0.490 |
| *M. musculus* | *H. sapiens* | 0.720 | 0.705 | 0.978 | 0.978 |
| *M. musculus* | *P. falciparum* | 0.181 | 0.134 | 0.196 | 0.126 |
| *M. musculus* | *S. cerevisiae* | 0.313 | 0.250 | 0.382 | 0.268 |
| *M. musculus* | *S. pombe* | 0.244 | 0.182 | 0.296 | 0.171 |
| *P. falciparum* | *A. thaliana* | 0.127 | 0.104 | 0.136 | 0.104 |
| *P. falciparum* | *C. elegans* | 0.126 | 0.113 | 0.132 | 0.114 |
| *P. falciparum* | *D. melanogaster* | 0.117 | 0.106 | 0.113 | 0.106 |
| *P. falciparum* | *H. sapiens* | 0.133 | 0.110 | 0.134 | 0.108 |
| *P. falciparum* | *M. musculus* | 0.128 | 0.116 | 0.129 | 0.120 |
| *P. falciparum* | *S. cerevisiae* | 0.132 | 0.111 | 0.143 | 0.114 |
| *P. falciparum* | *S. pombe* | 0.126 | 0.113 | 0.125 | 0.111 |
| *S. cerevisiae* | *A. thaliana* | 0.170 | 0.180 | 0.168 | 0.172 |
| *S. cerevisiae* | *C. elegans* | 0.184 | 0.198 | 0.201 | 0.226 |
| *S. cerevisiae* | *D. melanogaster* | 0.188 | 0.209 | 0.216 | 0.243 |
| *S. cerevisiae* | *H. sapiens* | 0.278 | 0.293 | 0.329 | 0.353 |
| *S. cerevisiae* | *M. musculus* | 0.160 | 0.168 | 0.166 | 0.176 |

| | | | | | |
|---|---|---|---|---|---|
| *S. cerevisiae* | *P. falciparum* | 0.163 | 0.158 | 0.195 | 0.185 |
| *S. cerevisiae* | *S. pombe* | 0.255 | 0.260 | 0.302 | 0.306 |
| *S. pombe* | *A. thaliana* | 0.223 | 0.231 | 0.217 | 0.224 |
| *S. pombe* | *C. elegans* | 0.205 | 0.212 | 0.220 | 0.229 |
| *S. pombe* | *D. melanogaster* | 0.228 | 0.246 | 0.258 | 0.285 |
| *S. pombe* | *H. sapiens* | 0.347 | 0.358 | 0.443 | 0.474 |
| *S. pombe* | *M. musculus* | 0.181 | 0.187 | 0.194 | 0.196 |
| *S. pombe* | *P. falciparum* | 0.140 | 0.127 | 0.147 | 0.130 |
| *S. pombe* | *S. cerevisiae* | 0.385 | 0.378 | 0.537 | 0.508 |

### *A.2 Cross-species tests for evolutionary accuracy*

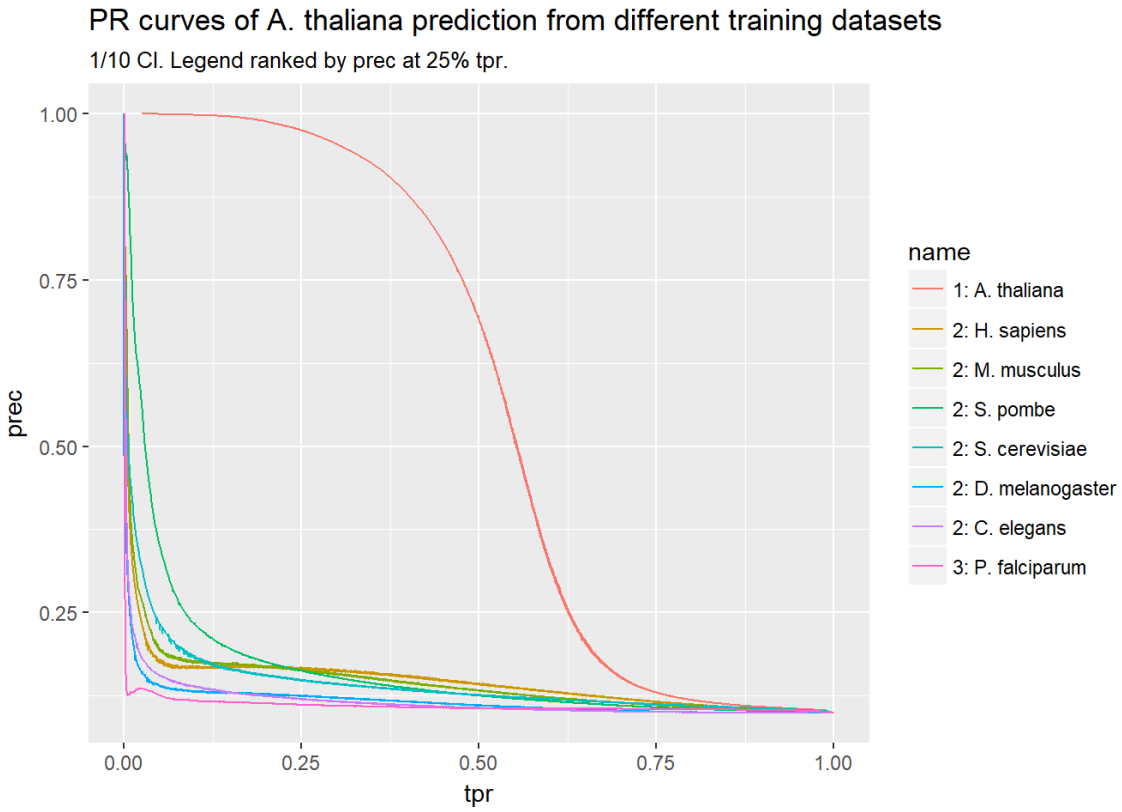



**Figure 53: P-R curve of predicting A. thaliana interactions from random sample of 2000 interactions from each training species at 10:1 CI (repeated 20 times, curves averaged).**

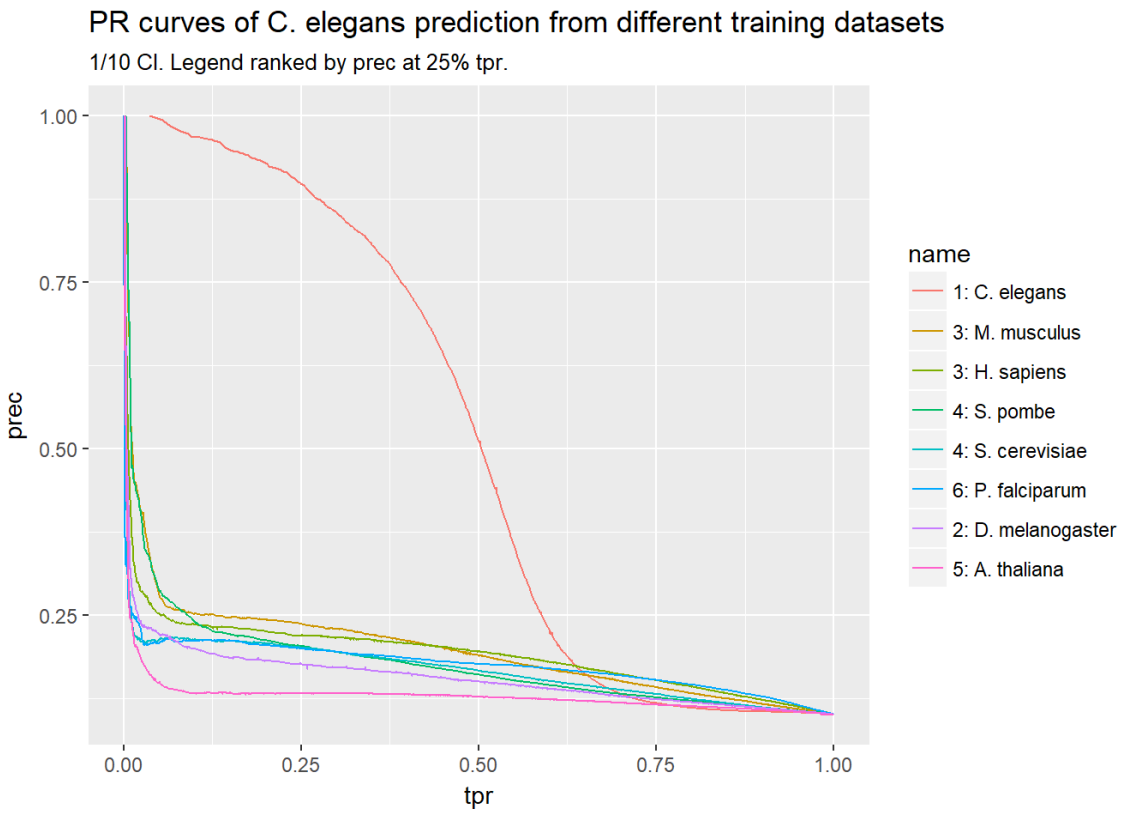



**Figure 54: P-R curve of predicting C. elegans interactions from random sample of 2000 interactions from each training species at 10:1 CI (repeated 20 times, curves averaged).**

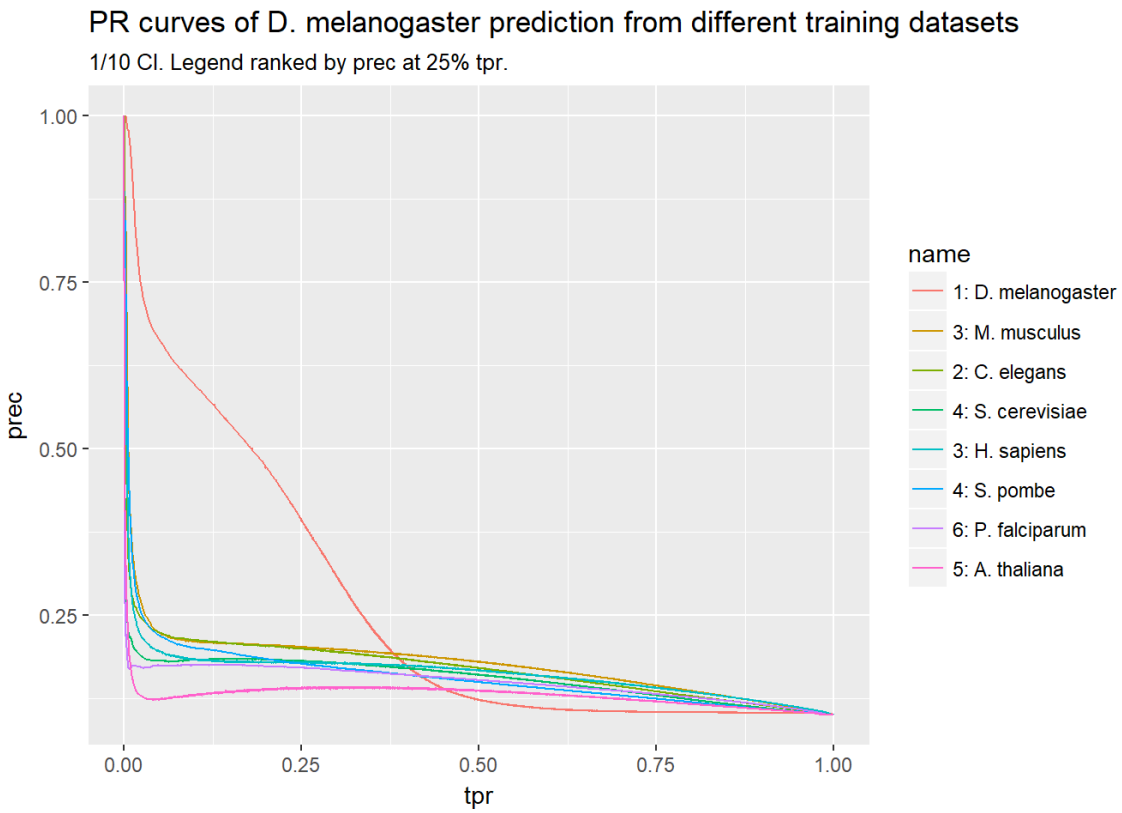



**Figure 55: P-R curve of predicting D. melanogaster interactions from random sample of 2000 interactions from each training species at 10:1 CI (repeated 20 times, curves averaged).**

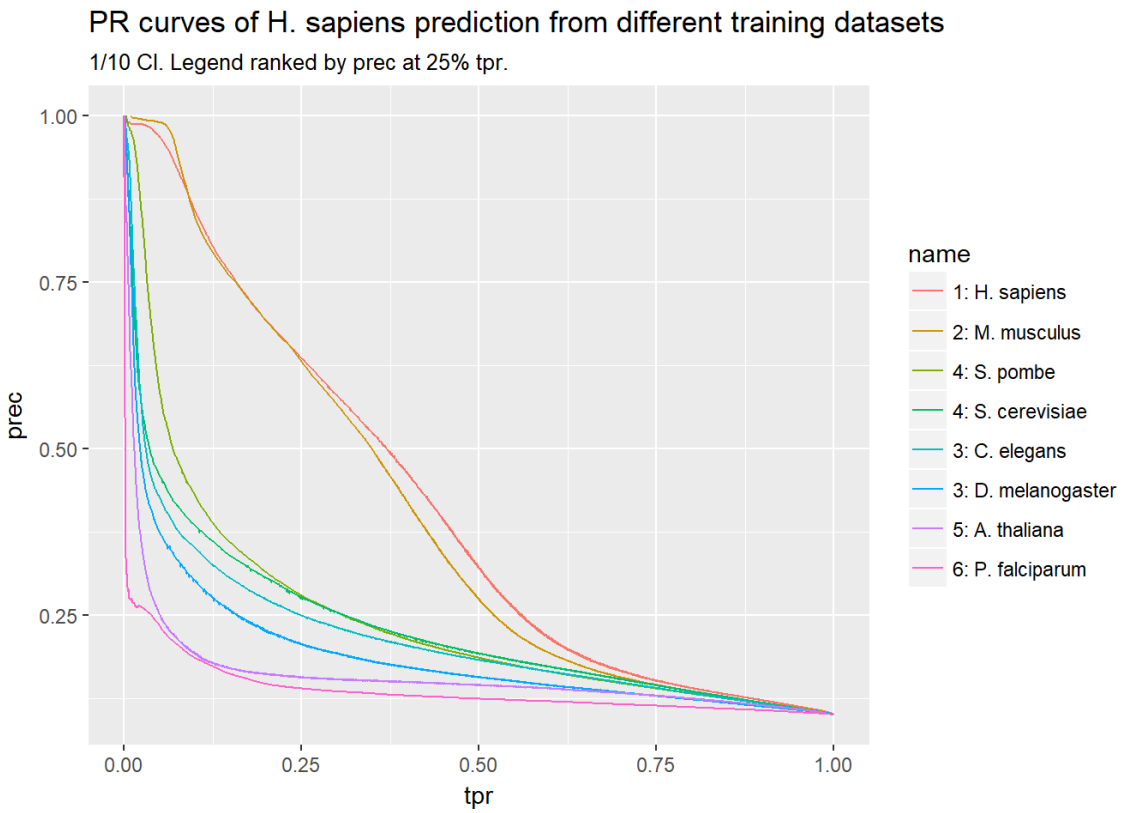



**Figure 56: P-R curve of predicting H. sapiens interactions from random sample of 2000 interactions from each training species at 10:1 CI (repeated 20 times, curves averaged).**

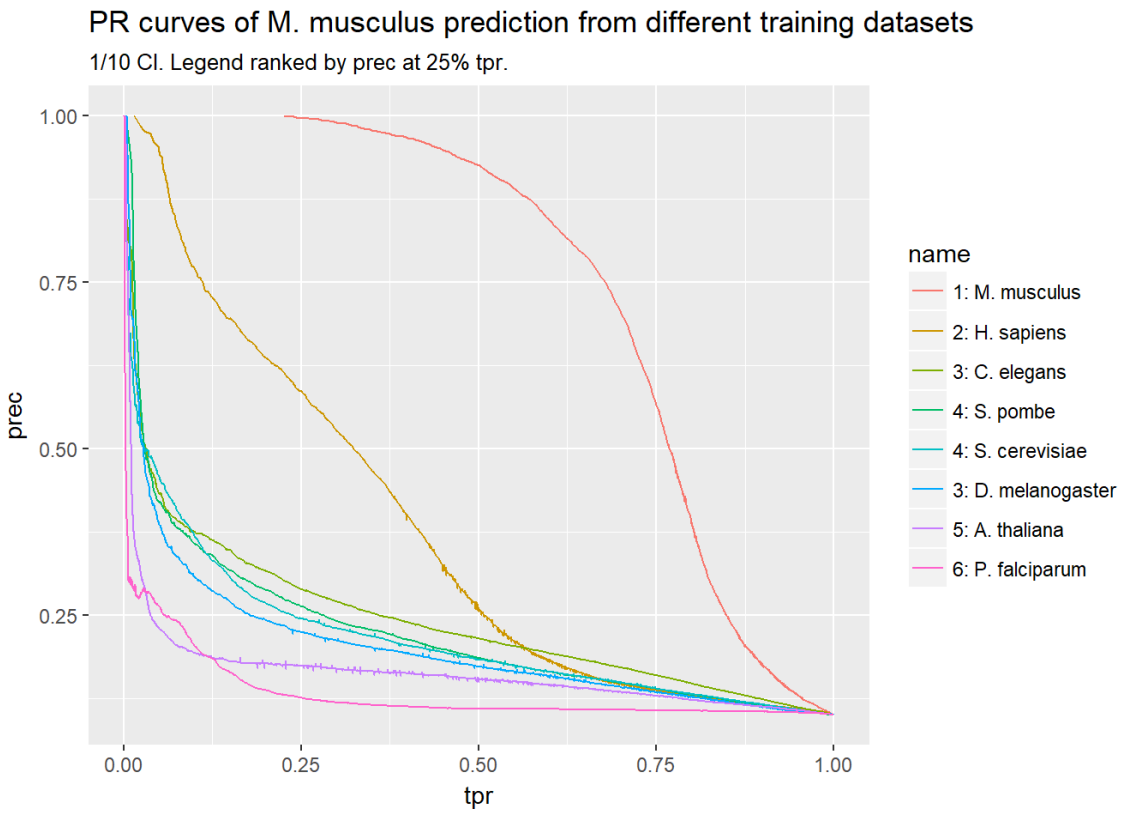



**Figure 57: P-R curve of predicting M. musculus interactions from random sample of 2000 interactions from each training species at 10:1 CI (repeated 20 times, curves averaged).**

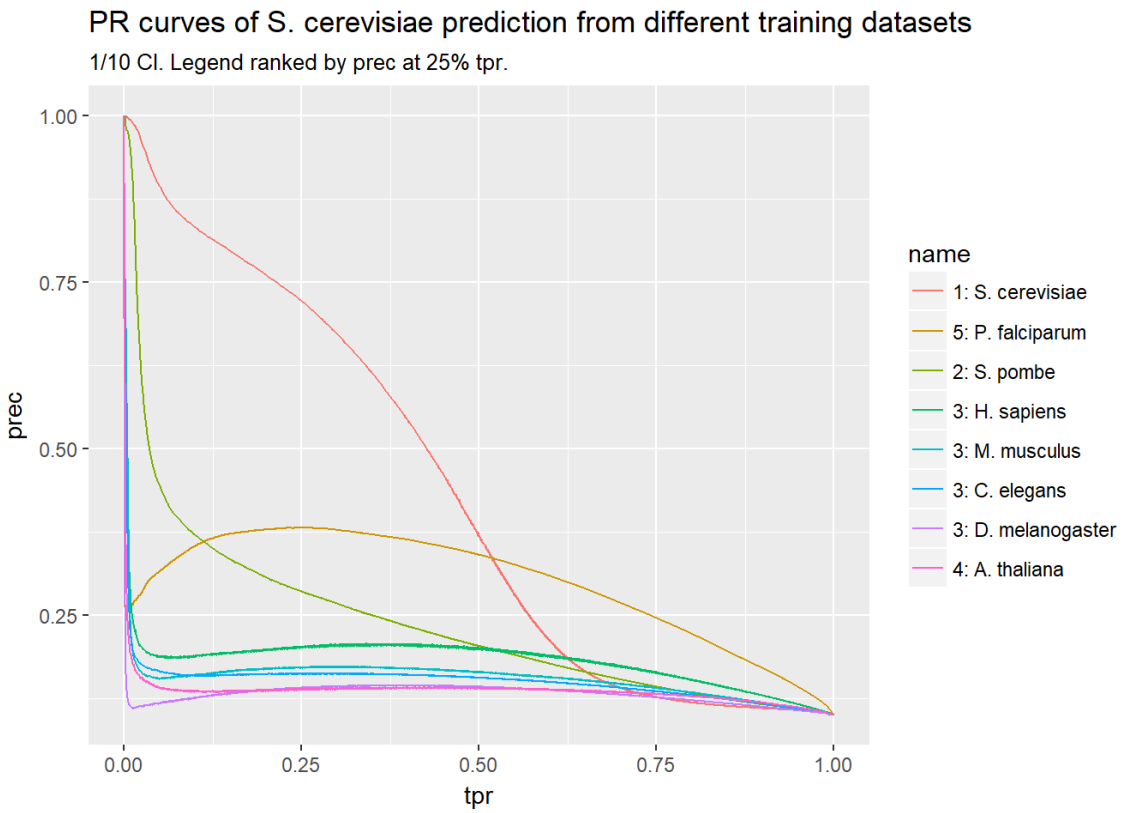



**Figure 58: P-R curve of predicting S. cerevisiae interactions from random sample of 2000 interactions from each training species at 10:1 CI (repeated 20 times, curves averaged).**

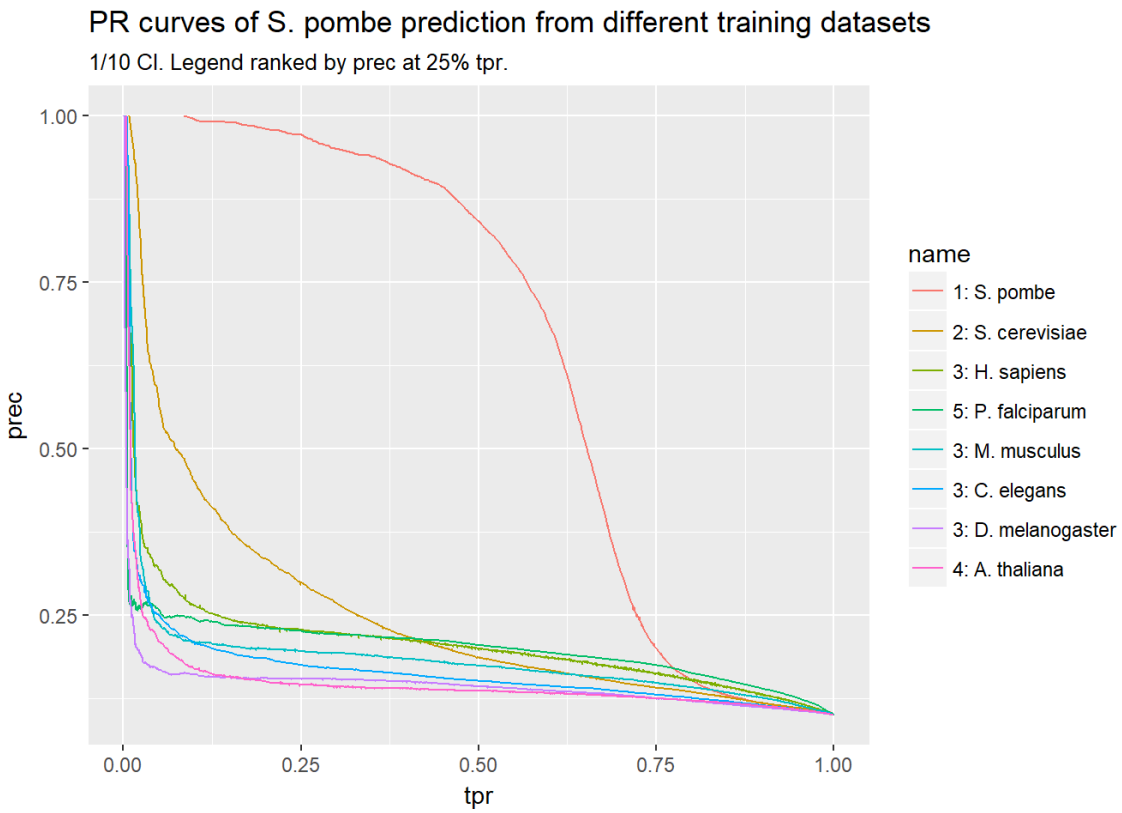



**Figure 59: P-R curve of predicting S. pombe interactions from random sample of 2000 interactions from each training species at 10:1 CI (repeated 20 times, curves averaged).**

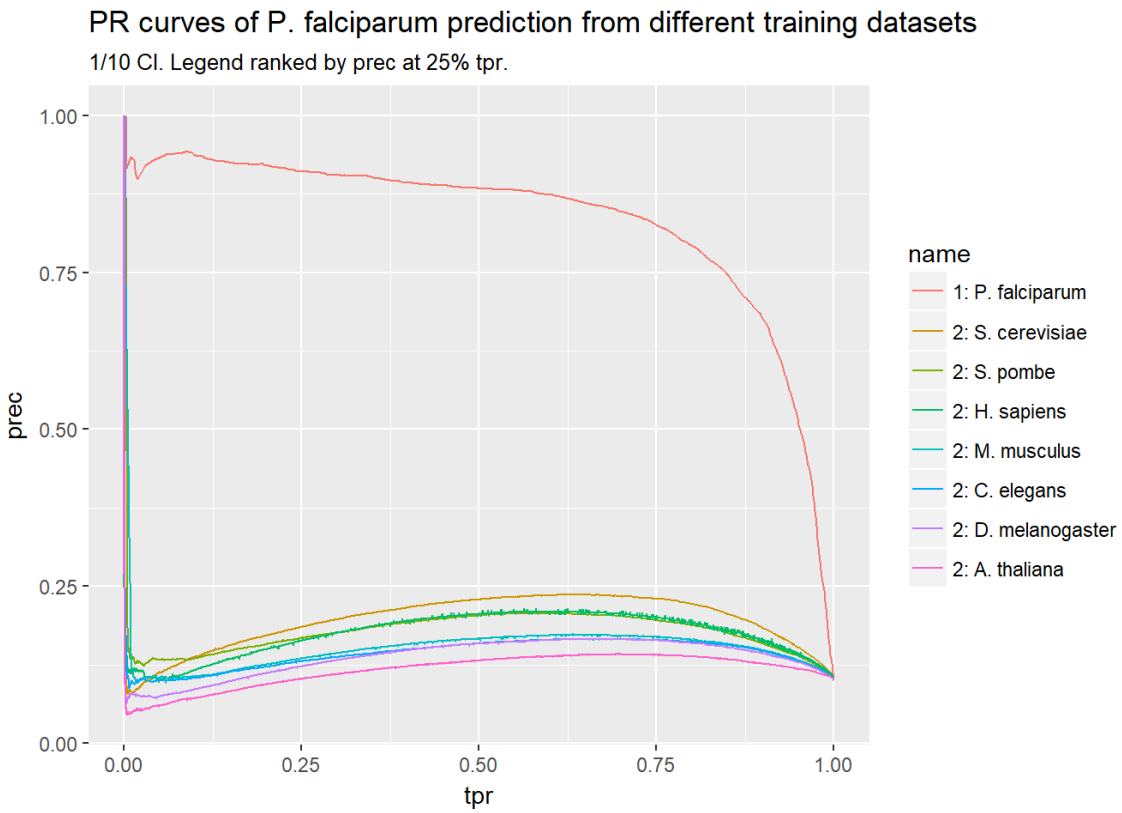



**Figure 60: P-R curve of predicting P. falciparum interactions from random sample of 2000 interactions from each training species at 10:1 CI (repeated 20 times, curves averaged).**

**Table 19: Full rankings for each performance metric of cross-species experiments (controlling for training set size).**

| | Expected Ranking | Evolutionary Ordering for correlation | AU-PRC ranks | Precision at 25% TPR ranks |
|---|---|---|---|---|
| *A. thaliana* | 1: A. thaliana | 1 | 1 | 1 |
| | 2: S. pombe | 2 | 2 | 2 |
| | 2: H. sapiens | 3 | 2 | 2 |
| | 2: M. musculus | 4 | 2 | 2 |
| | 2: S. cerevisiae | 5 | 2 | 2 |
| | 2: D. melanogaster | 6 | 2 | 2 |
| | 2: C. elegans | 7 | 2 | 2 |
| | 3: P. falciparum | 8 | 3 | 3 |
| *C. elegans* | 1: C. elegans | 1 | 1 | 1 |
| | 2: D. melanogaster | 2 | 3 | 3 |
| | 3: M. musculus | 3 | 3 | 3 |
| | 3: H. sapiens | 4 | 4 | 4 |
| | 4: S. pombe | 5 | 6 | 4 |
| | 4: S. cerevisiae | 6 | 4 | 6 |
| | 5: A. thaliana | 7 | 2 | 2 |
| | 6: P. falciparum | 8 | 5 | 5 |
| *D. melanogaster* | 1: D. melanogaster | 1 | 1 | 1 |
| | 2: C. elegans | 2 | 3 | 3 |
| | 3: M. musculus | 3 | 2 | 2 |
| | 3: H. sapiens | 4 | 3 | 4 |
| | 4: S. pombe | 5 | 4 | 3 |
| | 4: S. cerevisiae | 6 | 4 | 4 |
| | 5: A. thaliana | 7 | 6 | 6 |
| | 6: P. falciparum | 8 | 5 | 5 |
| *H. Sapiens* | 1: H. sapiens | 1 | 1 | 1 |
| | 2: M. musculus | 2 | 2 | 2 |
| | 3: C. elegans | 3 | 4 | 4 |
| | 3: D. melanogaster | 4 | 4 | 4 |
| | 4: S. pombe | 5 | 3 | 3 |
| | 4: S. cerevisiae | 6 | 3 | 3 |
| | 5: A. thaliana | 7 | 5 | 5 |
| | 6: P. falciparum | 8 | 6 | 6 |
| *M. musculus* | 1: M. musculus | 1 | 1 | 1 |
| | 2: H. sapiens | 2 | 2 | 2 |

| | | | | |
|---|---|---|---|---|
| | 3: C. elegans | 3 | 3 | 3 |
| | 3: D. melanogaster | 4 | 4 | 4 |
| | 4: S. pombe | 5 | 4 | 4 |
| | 4: S. cerevisiae | 6 | 3 | 3 |
| | 5: A. thaliana | 7 | 5 | 5 |
| | 6: P. falciparum | 8 | 6 | 6 |
| *S. cerevisiae* | 1: S. cerevisiae | 1 | 1 | 1 |
| | 2: S. pombe | 2 | 5 | 5 |
| | 3: H. sapiens | 3 | 2 | 2 |
| | 3: M. musculus | 4 | 3 | 3 |
| | 3: C. elegans | 5 | 3 | 3 |
| | 3: D. melanogaster | 6 | 3 | 3 |
| | 4: A. thaliana | 7 | 4 | 3 |
| | 5: P. falciparum | 8 | 3 | 4 |
| *S. pombe* | 1: S. pombe | 1 | 1 | 1 |
| | 2: S. cerevisiae | 2 | 2 | 2 |
| | 3: H. sapiens | 3 | 3 | 3 |
| | 3: M. musculus | 4 | 5 | 5 |
| | 3: C. elegans | 5 | 3 | 3 |
| | 3: D. melanogaster | 6 | 3 | 3 |
| | 4: A. thaliana | 7 | 4 | 3 |
| | 5: P. falciparum | 8 | 3 | 4 |
| *P. falciparum* | 1: P. falciparum | 1 | 1 | 1 |
| | 2: A. thaliana | 2 | 2 | 2 |
| | 2: S. pombe | 3 | 2 | 2 |
| | 2: H. sapiens | 4 | 2 | 2 |
| | 2: M. musculus | 5 | 2 | 2 |
| | 2: S. cerevisiae | 6 | 2 | 2 |
| | 2: D. melanogaster | 7 | 2 | 2 |
| | 2: C. elegans | 8 | 2 | 2 |