FCT
Fundação para a Ciência e a Tecnologia
MINISTÉRIO DA EDUCAÇÃO E CIÊNCIA

MIT Portugal

POPH
QUALIFICAR É CRESCER.

QREN
QUADRO DE REFERÊNCIA ESTRATÉGICO NACIONAL
PORTUGAL2007.2013

Governo da República Portuguesa

UNIÃO EUROPEIA
Fundo Social Europeu

Tiago Filipe Coelho Resende
Improving predictions in Metabolic Engineering problems by incorporating enzyme structural information
UMinho|2018

Universidade do Minho
Escola de Engenharia

Tiago Filipe Coelho Resende

Improving predictions in Metabolic Engineering problems by incorporating enzyme structural information

November 2018

Universidade do Minho
Escola de Engenharia

Tiago Filipe Coelho Resende

Improving predictions in Metabolic Engineering problems by incorporating enzyme structural information

PhD thesis in Bioengineering

This work was executed under the supervision of:
Professor Isabel Cristina A. Pereira da Rocha
Professor Cláudio Manuel Soares

November 2018

STATEMENT OF INTEGRITY

I hereby declare having conducted my thesis with integrity. I confirm that I have not used plagiarism or any form of falsification of results in the process of the thesis elaboration.

I further declare that I have fully acknowledged the Code of Ethical Conduct of the University of Minho.

University of Minho, November 27th 2018

Tiago Filipe Coelho Resende

Tiago Filipe Coelho Resende

Agradecimentos

Com a conclusão da etapa final no percurso do meu doutoramento gostaria de deixar um sentido agradecimento a todas as pessoas que fizeram parte da minha vida nestes últimos quatro anos.

Em primeiro lugar quero agradecer aos meus orientadores, Professora Isabel Rocha e Professor Cláudio Soares por todo o apoio e dedicação dados ao longo do doutoramento, pela paciência e disponibilidade demonstrada e pelo voto de confiança depositado que nos permitiu aprender e atingir objetivos que pareciam difíceis de atingir.

Agradeço ao programa MIT-Portugal pela oportunidade concedida e pela formação prestada nas diferentes universidades onde decorreram as aulas do ano curricular. Ao Centro de Engenharia Biológica da Universidade do Minho por ser a minha instituição de acolhimento e ao Instituto de Tecnologia Química e Biológica António Xavier (ITQB-NOVA) pela parceria na minha formação e disponibilidade para me acolher. Agradeço também à Fundação para a Ciência e Tecnologia pela atribuição da minha bolsa de doutoramento (SFRH/BD/96365/2013).

Aos colegas do BisBII pela companhia e apoio prestado ao longo destes anos sempre com disponibilidade e vontade de ajudar.

À minha família que sempre me motivou a seguir em frente e a acreditar que tudo é possível e pelo carinho sempre presente.

À D. Rosa, ao Nel e a Ana por me fazerem sentir sempre parte da família, por todo o apoio e por todos os momentos que sãosempre bem passados.

Aos amigos de sempre, por toda a amizade e conforto que proporcionam na minha vida e pelos momentos que passamos juntos e continuaremos a passar.

Por fim um agradecimento especial à Mariana, por todo o apoio e motivação ao longo deste doutoramento, mas ainda mais importante por todo o carinho e felicidade que partilhamos.

A todos, muito obrigado!

ABSTRACT

Systems biology foundations in this post-genomic era broadly rely upon well performed gene functional annotations. However, the experimental determination of a protein function is a laborious and expensive process, being its application unfeasible for the growing amount of sequences annotated. Current methodologies are based on computational tools that predict protein function over sequence similarity, but neglect a significant part of the putative proteins, ignore structural features and propagate annotation errors.

The present thesis focuses in the improvement of predictions in metabolic engineering problems by incorporating enzyme structural information. The uncertainty on the usage of NADP(H) or NAD(H) as co-factors was addressed due to the major impact in metabolic engineering applications these molecules have, severely affecting both predictions and strain design results. The molecular determinants for cofactor specificity were unveiled, using enzyme structural information from a representative dataset of enzymes present in Protein DataBank with NAD(P)(H) as cofactors, and support vector machines. The integration of homology modelling tools and a support vector machine predictive model allowed a process automation for predicting cofactor specificity. The resulting software was made available online in the form of a webserver. Cofactor prediction of a curated dataset of structurally uncharacterized enzymes was performed with success, validating the developed method.

The analysis and curation of the use of these cofactors in genome-scale metabolic models (GEM) was also performed through the cofactor prediction of sequences from 59 GEMs associated to reactions using NAD(P)(H). Results show some inconsistencies in GEM curation that can impair the correct simulation of the models, with an overall estimate of 28% of the genes implemented in the model being misclassified for cofactor specificity. The most recent GEM from *Saccharomyces cerevisiae*, Yeast 7.6, was corrected following the predictions performed and generally showed considerably improved results compared with the original version in the central metabolism, particularly for the flux in the pentose phosphate pathway. With the information on NAD(P)(H) cofactor specificity generated, a method was also developed and implemented to automatically predict the set of optimal gene mutations necessary for changing the NAD(P)(H) cofactor specificity of enzymes with unknown structure. Preliminary data shows promising results for the development of a tool for the automatic prediction of cofactor changing mutations, but experimental validation and further development are still a requirement.

RESUMO

A fundação da biologia de sistemas depende amplamente numa correta anotação de genes por homologia. Contudo, a determinação experimental da função de uma proteína é um processo moroso e dispendioso, sendo a sua aplicação inexequível para o crescente número de sequências anotadas. Metodologias atuais são baseadas em ferramentas computacionais que preveem a função de uma proteína de acordo com a similaridade da sua sequência com outras sequências caracterizadas, mas negligenciam uma parte significativa das proteínas putativas, assim como ignoram características estruturais e propagam erros de anotação.

A presente tese foca-se no melhoramento de previsões em problemas de engenharia metabólica através da incorporação de informação sobre estrutura enzimática. A incerteza da utilização de NAD(H) ou NADP(H) como cofatores foi abordada, devido ao enorme impacto em aplicações de engenharia metabólica que estas moléculas possuem. Os determinantes moleculares da especificidade de cofator foram desvendados através da utilização de informação estrutural presente numa base de dados representativa de enzimas com cofatores NAD(P)(H) presentes na PDBe da aplicação de *support vector machines* (SVM). A integração de ferramentas de modelação por homologia e um modelo preditivo de SVM permitiram a automatização do processo de previsão da especificidade de cofator. O *software* resultante foi disponibilizado online na forma de um *webserver*. A previsão dos cofatores de uma base de dados curada de enzimas sem informação estrutural foi efetuada com sucesso, validando o método desenvolvido.

A análise e curação do uso deste cofatores em modelos metabólicos à escala genómica (GEM) foi efetuada através da previsão de cofator das sequências associadas a NAD(P)(H) de 59 GEM. Os resultados mostraram algumas inconsistências na curação destes modelos que podem impedir a sua correta simulação, com cerca de 28% dos genes estando tendo o cofator mal classificado. O modelo mais recente de *S. cerevisiae*, Yeast 7.6, foi corrigido e simulado tendo mostrado resultados gerais melhorados no metabolismo central, particularmente na via da pentose fosfato, quando comparados com o modelo original. Com a informação sobre especificidade de cofatores gerada, um método foi desenvolvido e implementado para a previsão do conjunto ótimo de mutações necessárias para a alteração do cofator de enzimas sem informação estrutural. Os dados preliminares indicam resultados promissores para o desenvolvimento de uma ferramenta para a previsão automática de mutações para alteração de cofatores, embora seja ainda necessária validação experimental das hipóteses geradas.

TABLE OF CONTENTS

CHAPTER 1 Introduction 1
1.1 Context and motivation 3
1.2 Research aims 4
1.3 Outline of the thesis 5
1.4 References 6
CHAPTER 2 State of the art 9
2.1 Systems biology concepts 9
2.2 Enzymes and cofactors 12
2.3 Machine learning 22
2.4 Genome-scale metabolic models 25
2.5 References 27
Chapter 3 Unveiling the molecular determinants for NAD(P)(H) cofactor specificity using structural information 41
3.1 Introduction 43
3.2 Methods 46
3.2.1 Structure analysis and CNRPM generation 46
3.2.2 CNRPM dataset extraction 46
3.2.3 Machine learning 46
3.2.4 Comparative modeling for structure analysis 47
3.2.5 NiCofactor tool construction 47
3.2.6 NiCofactor result validation dataset 48
3.3 Results and discussion 48
3.3.1 Cofactor neighbor residue profile matrix (CNRPM) development 48
3.3.1.1 Building a comprehensive and representative CNRPM dataset 49
3.3.2 CNRPM dataset analysis and processing using Machine learning 49
3.3.3 NiCofactor cofactor specificity prediction tool development 53
3.3.3.1 Validation of NiCofactor cofactor specificity prediction tool using curated information 54
3.3.3.2 NiCofactor sensitivity analysis with case studies 56
3.3.3.3 NiCofactor tool usage for assessing cofactor engineering mutations impact in specificity 58
3.4 Conclusions 61
3.5 References 62
CHAPTER 4 Applying NiCofactor to the curation of NAD(P)(H) usage in Genome-scale metabolic models 69
4.1 Introduction 71
4.2 Methods 73
4.2.1 GEM download and Aminoacid sequence retrieval 73
4.2.2 NiCofactor aminoacid sequence processing 73
4.2.3 Simulations with GEMs 73

4.3 Results 73

4.3.1 Characterization of the GEM dataset 73

4.3.2 Model NAD(P)(H) cofactor usage analysis 78

4.3.3 NiCofactor cofactor prediction results 80

4.3.4 Comparison of genome-scale metabolic models from *Saccharomyces cerevisiae* 89

4.3.5 Analysis of *S. cerevisiae* Yeast 7.6 GEM 93

4.4 Conclusions 99

4.5 References 99

CHAPTER 5 Development of an *in silico* method for the conversion of NAD(P)(H) cofactor specificity in structurally uncharacterized enzymes 113

5.1 Introduction 115

5.2 Methods 117

5.2.1 Aminoacid residue sequences and protein structure templates 117

5.2.2 Cofactor specificity prediction 118

5.2.3 Evolutionary algorithm implementation 118

5.2.4 Protein structure visualization 119

5.3 Results and discussion 119

5.3.1 Identification of mutable residue positions for cofactor specificity reversal 119

5.3.2 Cofactor specificity reversal – Deterministic method 120

5.3.2.1 Implementation of the deterministic method for cofactor specificity conversion 122

5.3.3 Cofactor specificity reversal – Stochastic method 125

5.3.4 Case studies 126

5.3.4.1 PsXR – Pichia stipitis xylose reductase 127

5.3.4.2 GoXD –Gluconobacter oxydans xylitol dehydrogenase 130

5.3.4.3 PsXD –Pichia stipitis xylitol dehydrogenase 132

5.3.4.4 TiLD –Thermoactinomyces intermedius leucine dehydrogenase 134

5.4 Conclusions 136

5.5 References 138

CHAPTER 6 General conclusions and future work 141

APPENDIX 145

LIST OF FIGURES

Figure 2.1 - Representation of the nicotinamide adenine dinucleotide cofactors with NAD (left) and NADP (right). The only molecular difference between these two molecules is the phosphate in NADP, here shown in color. 16

Figure 2.2 - Depiction of the core of Rossmann fold, composed by two parallel β-strands separated by an α-helix. 17

Figure 2.3 - Graphical representation of a SVM. Composing the hyperplane representation are the features (blue for NAD(H) and white for NADP(H), along with the hyperplane separating both groups of features. 24

Figure 3.1 - Cofactor neighbor residue profile matrix generation. Starting with the whole structure, the tool pinpoints the location of the cofactor as a reference and proceeds to register the position of its atoms. For each cofactor atom, the tool investigates the surrounding residues and catalogues those within 6 Å. The complete process is performed automatically. The end result encompasses a matrix of interactions between each cofactor atom and neighboring residues. The presented enzyme structure depicts a Dihydropteridine Reductase bound to NAD$^+$ from Rat liver, with the EC 1.5.1.34. PDB id: 1DIR. 49

Figure 3.2 - Developed framework pipeline. The displayed planes depict the sequence of events necessary for cofactor prediction. Starting from the left top, a Fasta file composed by the target aminoacid sequences is inputted in the system where they are structurally modelled, analyzed and classified. 53

Figure 3.3 - SVM model validation graph. Plotted are all predictions performed (x axis) and the correspondent SVM model outputted prediction probability (y axis). Colors of the points refer correct (blue) and incorrect (red) predictions performed by the tool. Red lines delineate prediction probability of 80 and 95 percent, and evidence the amount of predictions encompassed in those probabilities. 55

Figure 3.4 - KGSADH II and KGSADH III aminoacid sequence alignment. KGSADH III (top) and KGSADH II (bottom) aminoacid sequence alignment. 56

Figure 4.1 - 1) Comparison between the total amount of genes included in each GEM (blue) with the number of genes in each GEM associated to reactions using NAD(P)(H) (red). 2) Comparison between the total amount of reactions included in each GEM (blue) with the number of reactions using NAD(P)(H) (red). 76

Figure 4.2 - Total amount of reactions using NAD(P)(H) (blue) and genes associated to reactions using NAD(P)(H) (red) in the selected 47 GEM, excluding the biomass equation and THD reactions (when occurring). 78

Figure 4.3 - Left top: Amount of Genes found to have a suitable structural template. In red are displayed the genes that have their structure characterized, or have a suitable structure template. In blue, the genes that do not have structural template. Left Bottom: From the total amount of genes with template, the amount of genes that had a prediction score equal or above 0.8 are called predicted genes. From the total amount of predicted genes, the amount that match with encoded reactions, the mismatches and the genes that 81 encode reactions with both cofactors. Right top: reactions using NAD(P)(H) as cofactors that have genes with template in their GPR. Green is the amount of reactions whose genes have structure template. Right bottom: in purple is amount of reactions encoded by the predicted genes. In red are the reactions that match the gene cofactor and in green those that do not.

Figure 4.4 - In the left is depicted a comparison between the total amount of genes and reactions included in each GEM. On the right, a comparison between the total amount of genes and reactions using NAD(P)(H) included in each GEM. 90

Figure 4.5 - Veen diagrams with the intersection of the common genes related to NAD(P)(H) contained in each model and also, in brackets, the intersection of the common predicted genes related to NAD(P)(H) contained in each model. 91

Figure 4.6 - Comparison between the six analyzed S. cerevisiae GEM. Left: Gene prediction comparison, with displayed data showing the amount of matching and mismatching predicted genes with reaction cofactor usage as well as genes present in reactions using both cofactors. Right: Reaction matches comparison displaying the amount of reactions, in each model, that match or mismatch the cofactor predicted. 92

Figure 4.7 - Comparison between the predictions of pFBA for the original and corrected models of Yeast 7.6 and also with experimental data. Original Yeast 7.6 (red), Corrected Yeast 7.6 (Green), Average of the 13C-MFA fully aerobic chemostat at a dilution rate of 0.1 h-1 (blue). Reactions: ACONT- aconitase, ACS- acetyl-CoA synthetase, ADH- alcohol dehydrogenase, AKGD- alpha-ketoglutarate dehydrogenase, ALD- aldehyde dehydrogenase, CSm- citrate synthase, FBA- fructose 1,6-bisphosphate aldolase, FUM- fumarase, G3PD1ir- glycerol-3-phosphate dehydrogenase, G3PT- glycerol-1-phosphatase, GAPD- glyceraldehyde-3-phosphate dehydrogenase, GHMT2r- serine hydroxymethyltransferase, GND- 6-phosphogluconate dehydrogenase, HEX- hexokinase, ICD- mitochondrial isocitrate dehydrogenase, MDH- mitochondrial malate dehydrogenase, MEmitochondrial malic enzyme, PDC- pyruvate decarboxylase, PDH- pyruvate dehydrogenase, PFKphosphofructokinase, PGI- phosphoglucose isomerase, PGK- 3-phosphoglycerate kinase, PGMTphosphoglucomutase, PP3- sum of the non-oxidative reactions of the pentose phosphate pathway producing glyceraldehyde-3-phosphate, PP6- sum of the non-oxidative reactions of the pentose phosphate pathway producing fructose-6-phosphate, PSP- phosphoserine phosphatase, PYC- pyruvate carboxylase, PYKpyruvate kinase, SUCD- succinate dehydrogenase, SUCOAS- succinyl-CoA ligase, THRS- threonine synthase, TPI- triose phosphate isomerase. Metabolites: 13dPG- 1,3-diphosphoglycerate, 3PG- 3-phosphoglycerate, AcCoA- acetyl-CoA, Akg- 2-oxoglutarate, Cit- citrate, DHAP- dihydroxyacetonephosphate, Fum- fumarate, G3P- glyceraldehyde-3-phosphate, Icit- isocitrate, Mal - L-malate, Oaaoxaloacetate, Ser- L-serine, Succ- succinate, SucCoa- succinyl-CoA. 97

Figure 5.1 – Process for the identification and selection of suitable mutable residue positions for cofactor specificity reversal. By combining the information on the target enzyme's CNRPM with the data in the SVM model, the sorting of best features for a specific enzyme is possible. By returning the features residue position in the sequence, a list of optimal mutable residues for cofactor specificity change is able to be generated. 120

Figure 5.2 - Cofactor atom-aminoacid interactions with highest impact on specificity. Each green circle represents an aminoacid, displayed in the single-letter format, which connects to each corresponding cofactor atom, forming the atom-aminoacid interaction with highest impact on specificity for each atom in both cofactors. 122

Figure 5.3 - Exemplification of the deterministic method implementation for cofactor specificity conversion. Given the ordered list of most influential residues for cofactor specificity and by consulting the point-mutation selection chart, the method selects and accumulates different mutations until the reverted cofactor specificity prediction is achieved. 124

Figure 5.4 – Evolutionary algorithm implemented in the development of the stochastic method for cofactor specificity reversal. During 100 generations, an optimization process is conducted, with multiple mutations being performed and analyzed with NiCofactor. Highest scoring mutants are retrieved and outputted as results. 126

Figure 5.5 - Depiction of the structure of PsXR Wild-type, with emphasis on cofactor binding-pocket, showcasing the wild-type with optimal residues for specificity reversal (top-left), literature experimental data point-mutations position (top-right), best prediction from stochastic method (bottom-left) and prediction from deterministic method (bottom-right). NAD+ is represented in blue, whereas selected residues are green. Labels indicate the original residue/sequence position/mutation. 129

Figure 5.6 - Depiction of the structure of GoXD Wild-type, with emphasis on cofactor binding-pocket, showcasing the wild-type with optimal residues for specificity reversal (top-left), literature experimental data point-mutations position (top-right), best prediction from stochastic method (bottom-left) and prediction from deterministic method (bottom-right). NADP+ is represented in light green, whereas selected residues are green. Labels indicate the original residue/sequence position/mutation. 131

Figure 5.7 - Depiction of the structure of PsXD Wild-type, with emphasis on cofactor binding-pocket, showcasing the wild-type with optimal residues for specificity reversal (top-left), literature experimental data point-mutations position (top-right), best prediction from stochastic method (bottom-left) and prediction from deterministic method (bottom-right). NADP+ is represented in light red, whereas selected residues are green. Labels indicate the original residue/sequence position/mutation. 133

Figure 5.8 - Depiction of the structure of TiLD Wild-type, with emphasis on cofactor binding-pocket, showcasing the wild-type with optimal residues for specificity reversal (top-left), literature experimental data point-mutations position (top-right), best prediction from stochastic method (bottom-left) and prediction from deterministic method (bottom-right). NADP+ is represented in yellow, whereas selected residues are green. Labels indicate the original residue/sequence position/mutation. 135

LIST OF TABLES

Table 2.1 - Summary of NAD(P)(H) cofactor engineering studies. Mutations are represented by the original aminoacid residue, in single letter code, followed by the residue position in the sequence and the mutant residue, also in a single letter coding format. Multiple mutations occurring in a single mutant are separated by a slash ('/'), while commas are used to separate individual mutants. 18

Table 2.2 - Depiction of the most important and available tools for protein structural prediction encompassing the three main methodologies, homology modelling, threading and *ab initio* as well as the format in which these tools are available. 22

Table 2.3 - Display of the different type of problems address in machine learning with the corresponding category. Applications and algorithms referent to these problems are also displayed. 23

Table 2.4 – Display of the most important tools available to perform genome-scale metabolic model reconstruction. A general description of the tool is also depicted. 26

Table 3.1 - Evaluation and performance parameters of the created SVM model. Accuracy, precision, MCC (Mathews correlation coefficient) and AUC ROC (area under the receiver operating characteristic curve) values (top) display the overall performance of the mode, indicating a well performing model. The Confusion matrix (bottom) evaluates sensitivity and specificity of the model. 50

Table 3.2 - SVM model feature weight distribution for NAD(H) (left) and NADP(H) (right). Feature weight is distributed in a decreasing order, starting from the heaviest. Columns depict the type of atom, residue (AA) and feature weight. Feature weights are divided into two sub columns for each cofactor. Colored cell in the atom columns represent the different areas composing the cofactor structure. 51

Table 3.3 - Cofactor specificity prediction. KGSADH II and KGSADH III from *Azospirillum brasiliense* cofactor specificity prediction analysis show the predicted cofactor and associated probability. ADH1 and ADH2 from *Geobacillus thermodenitrificans* NG80-2 cofactor specificity prediction analysis show the predicted cofactor and associated probability. Template information is also displayed with PDB ID and crystalized cofactor, as well as subject and template aminoacid sequence alignment identity percentage. 57

Table 3.4 - Wild-type enzymes with cofactor predictions above the probability threshold. Displayed are the organism, enzyme name, number of mutations implemented in the analyzed studies, wild-type cofactor specificity, predicted cofactor specificity and probability score. 59

Table 3.5 - Performed mutations and respective cofactor predictions. Displayed are also the organism and enzyme, along with mutant cofactor usage and the specification of the mutation performed. Mutations are represented by the original aminoacid, in single letter code, followed by the aminoacid position in the sequence and the mutant aminoacid, also in a single letter coding format. Multiple mutations occurring in a single mutant are separated by a slash ('/'). 60

Table 4.1 - Complete list of the retrieved and analyzed GEM, encompassing Model name, year of publication, organism name, taxa and reference. 74

Table 4.2 - GEM distribution of the amount of genes with template and predicted genes, along with the reactions from genes with template and the reactions from the predicted genes. 81

Table 4.3 - GEM distribution of the amount of predicted genes matching and mismatching reaction cofactor usage, as well as the amount of genes that encode reactions using both cofactors. 83

Table 4.4 - GEM distribution of the total amount of reactions with cofactor usage matching the predicted genes, as well as reactions mismatching the predicted genes cofactor and reactions with GPR composed of predicted genes with both cofactors. 87

Table 4.5 - Predicted genes with mismatching cofactor usage and the respective reactions. 93

Table 4.6 - Yeast 7.6_ Corrected model with the description of the performed corrections along with reaction identification and reaction name. 94

Table 5.1 - Point-mutation selecting chart. Each cofactor atom and corresponding molecular localization is represented in the column "cofactor atom". For each atom, the corresponding NAD(H) and NADP(H) specificity is represented in columns "NAD(H) specific" and "NADP(H) specific" respectively. 121

Table 5.2 - Mutations, cofactor predictions and prediction scores from literature experimental data, as well as from the implementation of both methods for reversing cofactor specificity in silico of PsXR. The deterministic method outputs only one mutant, while the stochastic method outputs five different mutants with the best found set of mutations for specificity reversal, according to the maximum candidate size allowed by the method, with the number on the gene name corresponding to the number of mutations selected. 128

Table 5.3 - Mutations, cofactor predictions and prediction scores from literature experimental data, as well as from the implementation of both methods for reversing cofactor specificity in silico of GoXD. The deterministic method outputs only one mutant, while the stochastic method outputs five different mutants with the best found set of mutations for specificity reversal, according to the maximum candidate size allowed by the method, with the number on the gene name corresponding to the number of mutations selected. 130

Table 5.4 - Mutations, cofactor predictions and prediction scores from literature experimental data, as well as from the implementation of both methods for reversing cofactor specificity in silico of PsXD. The deterministic method outputs only one mutant, while the stochastic method outputs five different mutants with the best found set of mutations for specificity reversal, according to the maximum candidate size allowed by the method, with the number on the gene name corresponding to the number of mutations selected. 132

Table 5.5 - Mutations, cofactor predictions and prediction scores from literature experimental data, as well as from the implementation of both methods for reversing 134 cofactor specificity in silico of TiLD. The deterministic method outputs only one mutant, while the stochastic method outputs five different mutants with the best found set of mutations for specificity reversal, according to the maximum candidate size allowed by the method, with the number on the gene name corresponding to the number of mutations selected.

Table A1 - The presented table encompasses all extracted feature weights from the developed SVM predictive model in chapter 3. 146

CHAPTER 1

Introduction

For millennia, human beings used the catalytic capabilities of microorganisms for their own advantage, unaware of the underlying biological principles and, therefore, unable to comprehend or control them. Within the past century a truly astonishing genomic revolution has befallen. With the discovery of the genome and genetic tools, we are now not only able to accurately describe biological processes, but also to manipulate genomes and metabolisms at our own will and desire. Nowadays, with metabolic engineering, bioinformatics and systems biology still rapidly evolving, our information processing proficiency has finally been bested by our data retrieving capacity. Nonetheless, hurdles are still to be surpassed.

In this post-genomic era, it has become expensive and timewise inefficient to experimentally characterize the functions of proteins, with current methodologies being based on computational tools that attribute protein function over sequence similarity. However, these methodologies (based on algorithms performing sequence homology searches) neglect a significant part of the putative proteins, ignore structural features and propagate annotation errors. There is therefore an urgent need for the development of new methodologies for predicting and identifying protein functions with a higher level of accuracy and using a broader approach. The work developed on this thesis focuses on the unveiling of molecular determinants for cofactor specificity and in the development of tools for more accurate predictions of protein functions and identification of cofactor specificities. Several approaches were undertaken, including the use of structure-based information, big-data and machine learning. The achieved results were applied in the correct identification of enzyme specificity for the seemingly equivalent cofactors nicotinamide adenine dinucleotide (NAD(H)) or nicotinamide adenine dinucleotide phosphate (NADP(H)) and in the analysis and curation of the use of these cofactors in genome-scale metabolic models. A method was also developed and implemented to automatically predict the set of optimal gene mutation necessary for achieving NAD(P)(H) cofactor specificity change.

1.1 Context and motivation

Advances in genome sequencing have allowed the number of sequenced organisms to dramatically increase in the last decade. In this post-genomic era, functional annotation has become a major concern of bioinformatics and systems biology, as the information required for the deployment of all metabolic processes taking place in cellular metabolism are encoded in the genome [1], [2].

As metabolic model reconstructions are becoming relevant tools for performing fundamental studies and for aiding drug discovery and bioprocess design, the impact of an accurate enzymatic function assignment from a genome sequence becomes evident [3]. Metabolic engineering and strain design endeavors also greatly benefit from well performed function annotations [4].

The experimental determination of a protein function is a laborious and expensive process, being its application unfeasible for the growing amount of sequences annotated [2], [5], [6]. The most common approach to function annotation, therefore, is based on the assumption that proteins with similar sequences perform similar functions [1], [2], [5]. This approach relies on algorithms using sequence alignment tools to analyze such similarities [6], [7]. However, errors spread easily when functional annotation is not done carefully due to overly unconstrained homology metrics [1], [7]. Also, there are many protein sequences without any previously characterized homologue, preventing function inference through similarity search [2]. Another problem of current functional annotation methodologies is the identification of the metabolic reactions that an enzyme catalyzes.

Catalytic activity of enzymes is classified by assigning an EC (Enzyme Commission) number, and for each EC number, there is a list of possible metabolic reactions. However, this reaction association might not be realistic, once there is insufficient curated information to confirm the catalytic activity of an enzyme in every possible reaction within an organism. This problem particularly affects Genome-scale metabolic model reconstruction due to the potential insertion and association of multiple misleading reactions [1], [2], [8]. In fact, it has been observed [9] that models reconstructed using the described methodologies may have limited use for metabolic engineering applications. Besides the existence of many gaps, the insertion of all the reactions associated to each EC number originates a metabolic model that has too many (artificial) degrees of freedom and that cannot be used for optimization [10]. Extensively curated models such as the ones reconstructed for *Escherichia coli* [11] and *Saccharomyces cerevisiae* [12] are not so affected by this problem, as available experimental characterization allows for an extensive refinement of predictions.

Although current technology in model development for biotechnology applications relies on the methodology described, tools have been developed that could be used to refine model building through existing methods that can be independent from similarity search [1], [2], [13]. Several computational approaches have been designed to improve the prediction of protein function, EC number classification and reaction association, using methods based on machine learning and training sets from databases with sequence/reaction profiling information [3], [6], [14]–[17].

Moreover, as the coverage of protein with structural characterization increases, this information can be used to generate homology models for many of these enzymes based on experimentally characterized templates, by analyzing and comparing them with the structural and functional data [18]–[21].

Further development of novel structure-based methods and the integration of machine learning methods in a practical platform might overcome existing flaws. Such platform shall represent a breakthrough in systems biology by largely increasing the accuracy of functional annotation and reducing annotation errors, being also potentially applied in the curation of genome-scale model reconstructions and in the process of developing new metabolic engineering strategies for strain design.

1.2 Research aims

The present PhD thesis focused on the study and comprehension of how protein function can be determined based on information other than relying solely on sequence similarity search in order to assist in the resolution of metabolic engineering problems. It was aimed to:

- Integrate information from different sources including:
    - Protein structural characterization;
    - Curated data on reaction cofactor specificity,
    - Experimental data;
    - Methods and datasets for the prediction of protein functions and structure.
- Develop new methods for understanding and predicting cofactor specificity of enzymes using NAD(P)(H);
- Develop a tool for performing cofactor predictions;
- Perform the curation of genome-scale metabolic models, using the developed tool;
- Develop a method for automatically predict mutants with reversed cofactor specificity.

1.3 Outline of the thesis

The research performed for this thesis is comprehended in six chapters. In the first chapter, an introduction and contextualization of the problems addressed in this thesis are presented. Chapters 2 to 5 further explore the problems and try several approaches for their solution. Chapter 6 encompasses the final research conclusions and recommends future approaches to be performed.

The four chapters exploring the research aims were organized as follows:

- In chapter 2, a comprehensive review on the state of the art of the extensive areas of expertise addressed in this thesis was performed. Literature and computational methods on genome functional annotation, metabolic engineering, enzyme structural characterization, machine learning and genome-scale metabolic models reconstruction and applications were analyzed.
- In chapter 3, the molecular determinants for NAD(P)(H) cofactor specificity were unveiled, using enzyme structural information. A comprehensive dataset of structures from enzymes using NAD(P)(H) as cofactors was built and processed using machine learning algorithms. The ensuing results were further analyzed to identify the responsible molecular factors for cofactor specificity. These findings where successfully applied to enzymes not structurally characterized, using comparative modelling and a protocol was developed to automatically predict cofactor specificity.
- In chapter 4, the developments aforementioned achieved were applied in the curation of genome-scale metabolic models from a wide range of microorganisms, through the correct characterization of reactions using the cofactors NAD(P)(H). Furthermore, the corrections suggested by the implemented curation were performed in the most recent model of *Saccharomyces cerevisiae*. Simulation results were then compared with the original model and *in vivo* experimental data for a performance assessment.
- In chapter 5, a method was developed for the automatic and high-throughput prediction of the optimal set of gene mutations required to obtain an efficient NAD(P)(H) cofactor specificity alteration. This was achieved through the analysis of the data outputted in the application of the machine learning algorithms used to predict cofactor specificity. The generated hypotheses were analyzed and applied *in silico* and the resulting enzyme mutants' cofactor specificity was predicted using the protocol developed in chapter 3.

1.4 References

[1] W. Tian and J. Skolnick, "How well is enzyme function conserved as a function of pairwise sequence identity?," *J. Mol. Biol.*, vol. 333, no. 4, pp. 863–882, 2003.

[2] D. Lee, O. Redfern, and C. Orengo, "Predicting protein function from sequence and structure," *Nat. Rev. Mol. Cell Biol.*, vol. 8, no. 12, pp. 995–1005, 2007.

[3] L. De Ferrari, S. Aitken, J. van Hemert, and I. Goryanin, "EnzML: Multi-label prediction of enzyme classes using InterPro signatures," *BMC Bioinformatics*, vol. 13, no. 1, pp. 1–12, 2012.

[4] V. G. Yadav, M. De Mey, C. Giaw Lim, P. Kumaran Ajikumar, and G. Stephanopoulos, "The future of metabolic engineering and synthetic biology: Towards a systematic practice," *Metab. Eng.*, vol. 14, no. 3, pp. 233–241, 2012.

[5] K. M. Borgwardt, C. S. Ong, S. Schonauer, S. V. N. Vishwanathan, A. J. Smola, and H.-P. Kriegel, "Protein function prediction via graph kernels," *Bioinformatics*, vol. 21, no. Suppl 1, pp. i47–i56, 2005.

[6] I. V. Tetko, I. V. Rodchenkov, M. C. Walter, T. Rattei, and H. W. Mewes, "Beyond the 'best' match: Machine learning annotation of protein sequences by integration of different sources of information," *Bioinformatics*, vol. 24, no. 5, pp. 621–628, 2008.

[7] B. Rost, "Enzyme function less conserved than anticipated," *J. Mol. Biol.*, vol. 318, no. 2, pp. 595–608, 2002.

[8] C. Z. Cai, L. Y. Han, Z. L. Ji, and Y. Z. Chen, "Enzyme family classification by support vector machines," *Proteins Struct. Funct. Bioinforma.*, vol. 55, no. 1, pp. 66–76, 2004.

[9] C. Zhang and Q. Hua, "Applications of genome-scale metabolic models in biotechnology and systems medicine," *Front. Physiol.*, vol. 6, no. JAN, pp. 1–8, 2016.

[10] I. Rocha, P. Maia, P. Evangelista, P. Vilaça, S. Soares, J. P. Pinto, J. Nielsen, K. R. Patil, E. C. Ferreira, and M. Rocha, "OptFlux: an open-source software platform for in silico metabolic engineering.," *BMC Syst. Biol.*, vol. 4, p. 45, 2010.

[11] J. D. Orth, T. M. Conrad, J. Na, J. a Lerman, H. Nam, A. M. Feist, and B. Ø. Palsson, "A

comprehensive genome-scale reconstruction of Escherichia coli metabolism—2011," *Mol. Syst. Biol.*, vol. 7, no. 535, pp. 1–9, 2011.

[12] H. W. Aung, S. A. Henry, and L. P. Walker, "Revising the Representation of Fatty Acid, Glycerolipid, and Glycerophospholipid Metabolism in the Consensus Model of Yeast Metabolism," *Ind. Biotechnol.*, vol. 9, no. 4, pp. 215–228, 2013.

[13] P. D. Dobson and A. J. Doig, "Predicting enzyme class from protein structure without alignments," *J. Mol. Biol.*, vol. 345, no. 1, pp. 187–199, 2005.

[14] J. L. Faulon, M. Misra, S. Martin, K. Sale, and R. Sapra, "Genome scale enzyme - Metabolite and drug - Target interaction predictions using the signature molecular descriptor," *Bioinformatics*, vol. 24, no. 2, pp. 225–233, 2008.

[15] V. Volpato, A. Adelfio, and G. Pollastri, "Accurate prediction of protein enzymatic class by N-to-1 Neural Networks," *BMC Bioinformatics*, vol. 14, no. Suppl 1, pp. 1–7, 2013.

[16] Y. Matsuta, M. Ito, and Y. Tohsato, "ECOH : An Enzyme Commission number predictor using mutual information and a support vector machine," *Bioinformatics*, vol. 29, no. 3, pp. 365–372, 2013.

[17] Y. Gao, B. Li, Y. Cai, K. Feng, Z. Li, and Y. Jiang, "Prediction of active sites of enzymes by maximum relevance minimum redundancy (mRMR) feature selection," *Mol. Biosyst.*, vol. 9, pp. 61–69, 2013.

[18] U. Pieper, B. M. Webb, G. Q. Dong, D. Schneidman-Duhovny, H. Fan, S. J. Kim, N. Khuri, Y. G. Spill, P. Weinkam, M. Hammel, J. A. Tainer, M. Nilges, and A. Sali, "ModBase, a database of annotated comparative protein structure models and associated resources," *Nucleic Acids Res.*, vol. 34, no. D1, pp. 291–295, 2006.

[19] A. Šali and T. L. Blundell, "Comparative protein modelling by satisfaction of spatial restraints," *Journal of Molecular Biology*, vol. 234, no. 3. pp. 779–815, 1993.

[20] B. Webb and A. Sali, *Comparative protein structure modeling using MODELLER*, vol. 2014. 2014.

[21] R. L. Chang, K. Andrews, D. Kim, Z. Li, A. Godzik, and B. O. Palsson, "Structural systems biology evaluation of metabolic thermotolerance in Escherichia coli," *Science (80-.).*, vol.

340, no. 6137, pp. 1220–1223, 2013.

CHAPTER 2

State of the art

2.1 Systems biology concepts

The use of microorganisms by humans for their catalytic capabilities in the preservation and enhancement of raw products dates back thousands of years, to the beginning of civilization itself. Countless products were created by altering their composition through the process of fermentation, despite the unconscious, and unknown, biological principles applied. It was not until the characterization of these organisms and their connection to food spoilage and alcoholic fermentation, by Pasteur, almost two centuries ago, that it was realized the importance and immense potential of their metabolism.

Upon realizing the potential application of their fermentative capabilities, efforts began to be made to understand and control such capacity. With a broader knowledge of the underlying mechanisms of microorganism metabolism, other fields of application also started thriving, with the development of food preservatives, antibiotics and synthetic compounds. Nonetheless, it was with the advent of DNA discovery [1], and its functional characterization [2], that the metabolism manipulation of microorganisms was finally unlocked, allowing the creation of the field of metabolic engineering [3], [4].

The main purpose of metabolic engineering is the highly efficient biosynthesis of added-value compounds, through the mutation of genes for the diversion of metabolic routes [5]. These mutations oblige the cell to redirect metabolic fluxes in the desired direction, generating microorganisms that are used as biocatalysts of biochemicals, biofuels and biopharmaceuticals, culminating in the secretion, or accumulation, of these target compounds [6].

In a first phase, this field operated using random mutagenesis, with experiments not depicting a clear view of the altered metabolic mechanisms [7]. This was due to the still very limited range of metabolic engineering resources and the concentrated efforts for gene modification, ignoring the overall metabolic portrait. However, the continuous development of mathematical and computational tools and sequencing technologies gave rise to a more rational metabolic engineering, combining modifications such as gene knockout, or overexpression, with predicted phenotype, allowing for a broader observation of the metabolic panorama [8]. The development of whole-genome sequencing, and consequent faster access to the information stored within the genome, promoted an easier introduction of targeted modifications in various organisms [9]. With the overwhelming and exponential amount of data being generated, new ways to handle the produced information were developed.

Using informatics and computational power, scientists developed new areas of expertise condensing knowledge from multiple fields and creating the field of bioinformatics and computational biology, pushing the beginning of the post-genomic era.

Along with the increase in the available information from genome sequencing, other types of experimental data also increased, supported by the development of novel high-throughput techniques. This allowed the combination of data from different sources in an attempt to produce a biological model accurately representing the strict relationship between phenotype and genotype. Systems biology comprehends the study of these complex systems, allowed by connecting genomic data with the biochemical reactions present in the metabolism of an organism [10]. Through the usage of computational and mathematical modeling, a quantitative description of the biological entities present in a cell is performed, predicting how the internal metabolism of the cells react to different input variables [11]–[13].

The genome of every organism stores all the information required for the correct deployment of every biochemical reaction and metabolic process taking place inside the cell, along with the interactions of the cell with the surrounding environment [14]. Information on gene expression, protein synthesis and regulatory and metabolic events are all coded in the genome, easily available due to the sequencing developments aforementioned. However, despite the abundance and accessibility of genomic sequence information, gene function characterization is not so trivial.

Genome annotation is the process of attributing a function to each putative gene in the genome. Gene function is regularly attributed through gene sequence similarity with a previously characterized gene. This process is performed using sequence alignment algorithms such us BLAST [15] and HMMER [16] that align the query sequence with characterized sequences from diverse databases, being the most popular, NCBI [17]. When two genes have a high sequence similarity they are called homologues, being the characterized function of one attributed to the other. When a homologue of a functionally uncharacterized gene is not found, the gene is annotated as a probable or hypothetical protein.

With the increasing amount of functional annotations being almost completely performed using homologue sequence similarity searches, an important issue arises. Since new genes are being functionally characterized with functions of genes that have gone through the same annotation process, there is little to no experimental evidence supporting new annotations, which leads to an increment in the introduction of errors and inconsistencies in the genome functional annotation [18], [19]. Moreover, with the increasing amount of automatic annotation pipelines being developed to cope with the exponential growth of sequenced genomes, this process becomes even more standardized, which largely expands the risks of poorly performed functional annotations [20]. The usage of sequence homology search in such pipelines, when not performed with extreme care, endlessly propagates annotation errors across all sequenced organisms, attributing outdated functions to new annotations and impairing the discovery of new gene functions [21].

Besides the information on gene functions retrieved from genome functional annotation, several other approaches and techniques are available to characterize the metabolism of an organism. Transcriptomics, proteomics, and metabolomics represent the follow up of genetic information in the central dogma of molecular biology and are among the most used and reliable approaches to characterize the metabolism of an organism. As the name suggests, transcriptomics performs the characterization of gene transcription, the genetic information being transcribed from the genome that will lead to the formation of proteins [22]. This characterization is performed by measuring mRNA expression and allows the identification of which genetic information is being used from the total amount of genomic information for any environmental condition. It also gives an exceptional insight of regulatory events occurring in the cell. Following gene transcription and translation, proteomics is used to study the expressed proteins in the organism. Using diverse experimental techniques, functional and structural analysis can be are performed, being the specific parameters of each protein function analyzed retrieved [23]. Metabolomics characterize, mainly using chromatographic techniques, the concentration of metabolites resulting from internal metabolic reactions, along with the metabolites secreted and consumed, depicting the organisms’ metabolic panorama [11], [24].

The combination of these approaches allow for a detailed understanding of the events occurring in an organism, from genotype to phenotype. Nonetheless, these methodologies are very expensive and laborious, being also relatively slow when compared to gene homology search, thus being only applied to specific cases and not inputted in a semiautomatic pipeline. For this reason, despite being precise, these approaches are considered inefficient processes.

2.2 Enzymes and cofactors

The metabolic reactions occurring in the metabolism of an organism, and encoded in its genome, are catalyzed by enzymes. This type of proteins possess catalytic properties, which accelerate biochemical reactions in the metabolism, being the responsible for the unique metabolic network present in each organism, allowing for a rapid and energy demanding metabolic event to occur [25].

Due to their outstanding qualities, enzymes are increasingly being used outside of the cellular environment, functioning as biocatalysts in the substitution of petroleum based synthetic chemistry for biofuel production and in the food and pharmaceutical industry [26], [27], to quote a few examples. Nonetheless, despite their importance, the molecular mechanisms of catalysis remain regularly quite elusive, demanding extensive analytic work for their full characterization and comprehension [28].

Being molecular recognition the principle responsible for the interactions occurring between enzyme and substrate, the accurate prediction and characterization of such interaction is of utmost importance for the correct determination of the set of reactions catalyzed by a given enzyme. The inaccurate prediction of these events impairs the correct depictions of an organism's metabolism, jeopardizing its use in metabolic engineering applications [29].

The association between annotated genes, enzymes and reactions is not a trivial task. In genomic functional annotation, the percentage of sequence identity and statistical score, are broadly used and established methods for assessing sequence similarity. Nonetheless, such methodology is not perfectly compatible when evaluating enzymatic function and the range of reactions catalyzed due to the fact that enzymatic function starts diverging early on in response to a small sequence similarity reduction [30].

In response to this, and to create a standardized process of enzyme function classification, the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology created the Enzyme Commission (EC) number as a hierarchical codification of enzyme classes. EC numbers represent a numerical code composed by 4 distinct elements. The first represents the main catabolic activity encoded, being divided in six categories: oxidoreductases (1); transferases (2); hydrolases (3); lyases (4); isomerases (5); and ligases (6). The remaining three elements detail the information on the reaction catalyzed, refining the classification to type of substrate accepted.

Being EC numbers the most used and accepted scheme for functional classification, such scheme is used as a mark in the process of binding gene sequence identity to precise function identification. However, this reaction association might not be realistic, once there is insufficient curated information to confirm the catalytic activity of an enzyme in every possible reaction within an organism, with often multiple reactions being attributed to a single EC number, leading to believe that an enzyme with such EC number would be capable of catabolizing all the reactions associated. Also, enzymes from different organisms, with the same EC number, can accept different types of substrates, depending on the molecular recognition set they have incorporated [14], [30], [31].

Enzyme function information can be retrieved from several databases with varying degrees of detail. Besides information based solely on sequence similarity and EC number, some databases also contain curated information on enzymes from individual organisms and characterized in diverse conditions. Uniprot (Universal Protein Resource Knowledgebase – www.uniprot.com) is one of the largest and most used databases covering protein functional information. It encompasses generally accurate and insightful information of protein function and is divided in two sections: swiss-prot, where curated information is stored; and TrEMBL, a database with automatic annotated records [32]. Brenda (braunschweig enzyme database – www.brenda-enzymes.org), is a database containing specifically and exclusively curated information on enzyme function. Its information is retrieved by dedicated curators from literature and contains various types of information, including kinetic parameters and enzyme interactions [33]. KEGG (Kyoto Encyclopedia of Genes and Genomes – www.kegg.jp) is one of the major containers of information on genes, enzymes, metabolites, reactions and pathways. Despite its importance and comprehensive repository on reaction/enzyme information, this database is not fully curated and should be inspected with care to avoid the implementation of inconsistencies in enzyme information [34].

Although current technology on enzyme functional characterization for biotechnology applications mainly relies on the described methodologies, several computational approaches have been designed to improve the prediction of protein function, EC number classification and reaction association, using methods based on Machine Learning and training sets from databases with sequence/reaction profiling information [29], [35]–[38].

One of the major contributions for enzyme reaction complexity, other than enzyme structural and molecular conformation, lies in the fact that, besides the catabolized substrate and consequent product release, enzymatic reactions encompass other reaction components that contribute to a well performed catalysis. Among the most prominent reaction participants are cofactors; metabolic intermediaries that transfer chemical groups between different metabolites, allowing for the reaction to occur. Cofactors are organic compounds or metal ions, and are often required for enzyme activity. Some of these low molecular weight entities are responsible for energy transfer in the cell, having the role of redox carriers for catabolic reactions, being essential in many reactions [39].

Some cofactors are closely bound to the enzyme and are usually self-regenerating, such as pyridoxal phosphate, biotin or flavins; while other cofactors, such us pyridine dinucleotides: nicotinamide adenine dinucleotide (NAD(H)) and nicotinamide adenine dinucleotide phosphate (NADP(H)), act as functional group transfer agents, being therefore consumed at the same rate of substrate consumption [40].

The essential cofactors for a correct cell metabolism are NAD(H), NADP(H) , S-adenosyl-methionine, flavin adenine dinucleotide, pyridoxal 5-phosphate, coenzyme A, thiamin diphosphate and flavin mononucleotide [41]. Of these, cofactors NAD(H) and NADP(H), are the most widely used and of great importance, being extensively examined for chemical processing applications [42].

An example of their scope is the fact that NAD(P)(H) are the cofactors responsible for catalyzing the transfer of electrons between molecules in the vast majority of Oxidoreductases, the largest group of enzymes in metabolism. In fact, according to Swissprot, the curated branch of Uniprot, from the total amount of ~556.000 reviewed protein sequences deposited in March 2018, approximately 35.000 (6.3%) are annotated as enzymes having NAD(H) or NADP(H) as ligands. A thorough comprehension of the molecular events taking place between enzymes and these cofactors is, thus, essential to a complete and accurate metabolic rebuilding of the individual reaction network belonging to every organism.

Due to the heavy regulation, by cells, of the levels of reduced and oxidized metabolic pools of NAD(P)(H), understanding cofactor specificity is of great value, in order to evaluate the routes available for the metabolic engineering of biological pathways and systems that involve these enzymes [39]. This regulation is required in order to maintain cofactor balance and availability, with some examples of common metabolic engineering hurdles including: butanol conversion yield in the production of biofuels [43]; xylose to ethanol fermentation yield, due to the formation of xylitol in a pathway using both cofactors [44]–[50]; cofactor regeneration requirement when used in oxidation reactions catalyzed by cytochrome P450 [51]–[53] and in the production of chiral chemical intermediates with pharmaceutical application, such us 4-hydroxy-2-butanone, that also require cofactor regeneration [54], [55].

Despite their resemblance, the role of each cofactor in metabolism is distinct, with the general conception that NAD(H) participates in the cell's catabolic processes, releasing energy from molecule breakdown, while NADP(H) participates in the anabolic processes, where large molecules are synthesized [56], [57]. In fact, NADP(H) is the reducing cofactor in many pathways, being regenerated in pentose phosphate pathway (PPP), as well as in isocitrate dehydrogenease and malic enzyme reactions. To increase the availability of this cofactor, efforts have been made to increase the consumption of NAD^+ while producing $NADP^+$, with the overexpression of NAD^+ Kinase, an enzyme that consumes ATP to convert NAD^+ in $NADP^+$. These efforts were successful in the increase of NADPH concentration in the cytosol, with NADH concentration decreasing. This change in cofactor pools had a positive effect in ethanol and acetate production in yeast during anaerobic growth on glucose, while xylitol production increased during anaerobic growth on xylose [58]–[60].

These studies showed the intrinsic relation between NAD(H) and NADP(H), and highlighted the importance of extensively characterizing the enzymes using these cofactors, as well as the molecular and structural determinants behind cofactor specificity.

Nonetheless the increasingly fast development of systems biology and metabolic engineering, cofactor specificity determination is still a hurdle, with elusive principles and molecular mechanisms that impair a common strategy, transversal to all enzymes using these cofactors. Despite the participation in a wide range of reactions, turning them in great metabolic targets, the fact that NAD(H) and NADP(H) are two very similar cofactors also turns their utilization in a challenging task, as NAD(P)(H) using enzymes often do not share significant sequence identity and cannot be easily detected by sequence homology [61].

The only difference in these cofactors, as their name suggests, is the presence of a phosphate group in the vicinity of the 2’ hydroxyl of the adenosine ribose, in NADP(H) as depicted in figure 2.1.

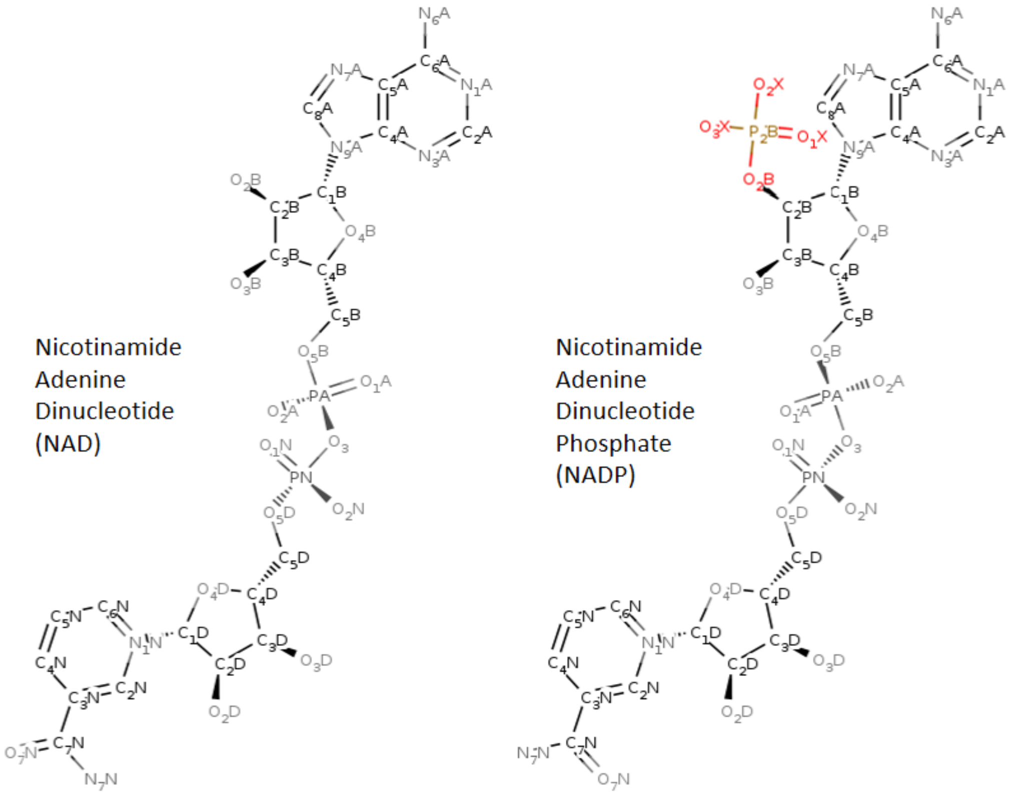


Figure 2.1 - Representation of the nicotinamide adenine dinucleotide cofactors with NAD (left) and NADP (right). The only molecular difference between these two molecules is the phosphate in NADP, here shown in color.

Along with their similarity, despite their distinctive existences, oxidation and reduction processes of these molecules are carried in the exact same location, in the nicotinamide moiety. This trait puts the only difference between both cofactors almost in the opposite side of the structure, relatively to where the chemical transition is occurring. This fact indicates that their difference has no effect in the type of reactions catalyzed, nonetheless, most enzymes exhibits a strict specificity for one of the cofactor [62]. This specificity allows the cell to regulate different metabolic pathways only by managing the concentration of these cofactors.

Despite efforts in predicting cofactor specificity, recent advances mainly focus in small protein family groups or specific traits, such us structural motifs [63], lacking approaches able to accurately perform prediction with a large portion of NAD(P)(H) binding enzymes. Also, few methods exist to try to determine the molecular determinants responsible for cofactor specificity and, when these exist, are specific for certain enzyme families, such as Ketol-acid reductoisomerases (KARI) [64].

Several efforts have also investigated the nucleotide binding domain of NAD(P)(H) binding enzymes, in search for differentiating traits in structural motifs [65]. The most common structural motif found in NAD(P)(H) binding enzymes is the Rossmann fold [66], but there are also less common and not specific motifs, such us TIM-barrel [67], the dihydroquinoate synthase-like and the FAD/NAD$^+$ binding folds [68].

The Rossmann fold is a common structural motif found in many nucleotide-binding proteins. At its core are two parallel β-strands separated by an α-helix (βαβ motif) as depicted in figure 2.2. A tight loop at the end of the first β-strand makes direct contact with the cofactor. Some consensus sequences have been proposed for this fold, with the first being the phosphate-binding sequence Gly-X$_{1-2}$-Gly-X-X-Gly. However, its short sequence was not reliable as a search motif. Later, with the input of other research [69], an extended Gly-X$_{1-2}$-Gly-X-X-Gly-X-X-X-[Gly/Ala] motif was proposed as an indicator of Rossmann folds that bind FAD or NAD(P), nonetheless some inconsistencies were also detected in the last residue, being considered variable [61].

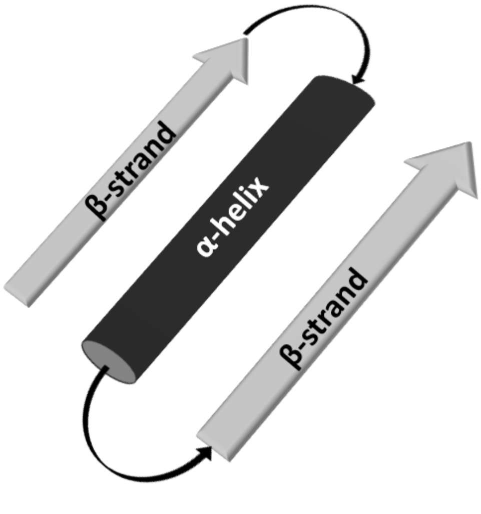


Figure 2.2 – Depiction of the core of Rossmann fold, composed by two parallel β-strands separated by an α-helix.

Notwithstanding these findings, structural studies have found that the number and spacing of conserved Glycine residues varies. In fact, it was concluded that in nearly three-quarters of the analyzed dataset at least twelve distinct locally conserved structural motifs for binding NAD(P)(H) were represented, having the remaining dataset no distinguishable motifs [70].

Some studies involving ketol-acid reductoisomerases have also shown that the presence of acidic residues at normally conserved phosphate binding positions are potential candidates of enzymes preferring NAD(H) [64]. Moreover, an early study concluded that the only conserved structural feature in NADP(H) complexes is an Arginine residue present near the adenine moiety, interacting with the phosphomonoester through hydrogen bonds. For NAD(H) it was determined that the only clue for specificity identification was the presence of an Aspartate (or Glutamate) residue able to chelate the diol group of the ribose near the adenine. However, frequently these structural features are not present in structures bound to NAD(H) or NADP(H), hindering their correct characterization, and besides, structural information on the subject enzyme is required, which regularly is not the case [62].

Regardless of the difficulties in correctly characterizing enzyme cofactor specificity from sequence or structure, many cases have been accomplished in metabolic engineering, concerning the alteration of NAD(P)(H) cofactor specificity. These efforts have, nonetheless, been accomplished with partial structural data to guide mutagenesis strategies. Table 2.1 depicts the most recent list of enzyme redesigns for altered NAD(P)(H) cofactor specificity using site-directed mutagenesis, as presented by Cui and coworkers [31] and extending from previous compilations [71], [72].

Table 2.1 - Summary of NAD(P)(H) cofactor engineering studies. Mutations are represented by the original aminoacid residue, in single letter code, followed by the residue position in the sequence and the mutant residue, also in a single letter coding format. Multiple mutations occurring in a single mutant are separated by a slash ('/'), while commas are used to separate individual mutants.

Source	Enzyme	Specificity	Mutation(s)	Ref
Thermus flavus	Malate dehydrogenase	NADH -> NADPH	E41-K47 loop	[73]
Bacillus stearothermophilus	L-lactate dehydrogenase	NADH -> NADPH	I51K/D52S	[74]
Rattus norvegicus	Cytochrome b5 reductase	NADH -> NADPH	D239T	[71]
Thermus thermophilus	β-isopropylmalate dehydrogenase	NADH -> NADPH	D236R/D289K/ I290A/A296V/G337Y	[75]
Drosophila melanogaster	Alcohol dehydrogenase	NAD$^+$ -> NADP$^+$	D38Q	[76]
Lactobacillus delbrueckii subsp. bulgaricus	D-lactate dehydrogenase	NAD$^+$ -> NADP$^+$	D175A	[77]
Bacillus stearothermophilus	Glyceraldehyde-3-phosphate dehydrogenase	NAD$^+$ -> NADP$^+$	D32A/L187A/P188S	[78]
Gluconobacter oxydans	Xylitol dehydrogenase	NAD$^+$ -> NADP$^+$	D38S/M39R	[79]
Homo sapiens	Human mitochondrial NAD(P)-dependent malic enzyme	NAD$^+$ -> NADP$^+$	Q362K	[80]
Pichia stipitis	Xylitol dehydrogenase	NAD$^+$ -> NADP$^+$	D207A/I208R/F209S /N211R	[81]
Pseudomonas stutzeri	Phosphite dehydrogenase	NAD$^+$ -> NADP$^+$	E175A/A176R	[82]
Saccharomyces cerevisiae	Formate dehydrogenase	NAD$^+$ -> NADP$^+$	D196A/Y197R	[83]
Thermus thermophilus	Isopropylmalate dehydrogenase	NAD$^+$ -> NADP$^+$	S226R/D278K/I279Y /A285V/P324T/	[84]

			P325Y/G328E/G329R/S330L	
Tramitichromis intermedius	Leucine dehydrogenase	NAD$^+$ -> NADP$^+$	D203A/I204R/D210R	[85]
Pseudomonas mevalonii	HMG-CoA reductase	NADPH -> NADH	D146A + L148R	[86]
Candida tenuis	Xylose reductase	NADPH -> NADH	K274R, K274G, K274M, S275A, N276D, R280H, K274R/N276D	[87], [88]
Corynebacterium	2,5-diketo-D-gluconic acid	NADPH -> NADH	K232G, R235G, R238H, F22Y/RS233T/R235E/A272G	[89], [90]
Escherichia coli	Glutathione reductase	NADPH -> NADH	A179G/A183G/V197E/R198M/K199F/H200D/R204P	[91]
Escherichia coli	Ketol acid reductoisomerase	NADPH -> NADH	R68D, K69L, K75V, R76D	[92]
Neurospora crassa	Nitrate reductase	NADPH -> NADH	S920D/R932S	[93]
Pichia stipitis	Xylose reductase	NADPH -> NADH	K270M,K270S/S271G/N272P/R276F	[94], [95]
Pseudomonas fluorescens	p-hydroxybenzoate hydroxylase	NADPH -> NADH	R33S/Q34R/P35R/D36A/Y37E	[96]
Rattus norvegicus	Cytochrome p450 reductase	NADPH -> NADH	S596D	[97]
Saccharomyces cerevisiae	17b-hydroxysteroid dehydrogenase	NADPH -> NADH	Y49D	[98]
Sinorhizobium Morelense	1,5-anhydro-D-fructose Reductase	NADPH -> NADH	A13G/S33D	[99]
Anabaena. sp. (strain PCC 7119)	Ferredoxin: NADP$^+$ reductase	NADP$^+$ -> NAD$^+$	S223D	[100]
Escherichia coli	Isocitrate dehydrogenase	NADP$^+$ -> NAD$^+$	C201I/C332Y/K344D/Y345I/V351A/Y391K/R395S	[101]
Thermus thermophilus	Isocitrate dehydrogenase	NADP$^+$ -> NAD$^+$	K283D/Y284I/N287G/V288I/I290A	[102]
Vibrio harveyi	Aldehyde dehydrogenase	NADP$^+$ -> NAD$^+$	T175D, T175E, T175S, T175N,T175Q	[103]

The implemented studies show that cofactor specificity change can balance cofactor availability, increasing pathway yields for multiple products. However, the fact that multiple simultaneous mutations had to be performed in order to effectively change cofactor specificity shows that there is a non-additive effect in the mutations performed, possibly due to the difference in the cofactor's structure [104], hindering even more an already challenging task.

The precise order of each amino acid residue position determines how a protein structure folds into its final conformation. The folded state of a protein reveals its native optimal conformation, and despite being the result of a natural and almost instantaneous process, protein folding is an extremely complex process to model, due to its large conformational search space [105]. In fact, Levinthal's paradox speculated that finding the native folded state of a protein by random search should take more time than the age of the universe [106]. This intricate folding conformation is the main responsible for enzymes biocatalytic capabilities, facilitating the interactions between different substrates and cofactors and promoting biochemical reactions. An enzyme's structural conformation not only determines the spatial environment for the reaction to occur, but also defines the molecular recognition apparatus responsible for molecular binding.

With substrate and cofactor specificity elusively imprinted in the structural conformation of each enzyme, the need to accurately and consistently predict these interactions is vital, with obvious direct application in genome functional annotation, systems biology and metabolic engineering [29].

The structural characterization process is performed by determining the exact localization of each atom in the protein's molecular structure relatively to every other atom in the molecule. The most widely used experimental methods to accomplish this characterization are X-ray crystallography and Nuclear Magnetic Ressonance (NMR). In X-ray crystallography, the protein molecule is purified and crystallised. An intense X-ray bean is them directed to it and diffracted into specific patterns that are posteriorly interpreted to generate a structure. In NMR, the subject protein is also purified and a solution of it is placed in a strong magnetic field, where it is probed with radio waves. The resulting resonances can be deconvoluted into structural data that allows the determination of the structure.

Once characterized, the spatial coordinates of each atom composing the protein's molecular structure are stored in a file along with coordinates of other types of molecules, such us cofactors, ions and water. These files can then be transferred to online repositories, such us Protein Data Bank (PDB – www.rcsb.org), the largest online database of protein structural information, with over 138 thousand available structures in March 2018 [107]. Despite their utility and accuracy, these experimental methods are expensive and time consuming, becoming inefficient with the increasing amount of genomic information generated. To overcome this hurdle, several protein structure modelling approaches are available, allowing the prediction of a protein's structure without direct experimental characterization.

Protein structure prediction methods are mainly divided in three different approaches, depending on the existence, and similarity, of template proteins with known structure. When the target aminoacid sequence does not have any available sequence homologue with known structure, protein structure can be modelled using *ab initio* (latin for: from the beginning) prediction methods. These methods attempt to predict the native state of the protein structure from its amino acid residue sequence, without templates [105].

With the general assumption that proteins fold to a state of minimal energy conformation, *ab initio* approaches can use molecular dynamics simulation (MD) to simulate the conformational behavior of proteins. However, the application of these methods, even for small residue proteins, is computationally very expensive, with the normal limit of simulations being microseconds, which is still far from the timescale of folding of many proteins [108].

When a template structure is identified, template-based methods, such us comparative modelling, are implemented.

In comparative modelling, the tridimensional structure of the target protein is modelled based on the structure of the template protein(s). For that, proteins with known structure and high sequence homology to the target sequence are sought to be used as templates. The sequence of both target and template are aligned, and the target structure is modelled from the coordinates of the template residues through, for example, the satisfaction of spatial restraints [109]. These restrains are retrieved from the assumed similar special information between aligned residues in the template and the target structures. An optimization process for minimizing all restrain violations is afterwards applied. Structural similarity is assume when two proteins share a sequence identity equal or above 30% [110].

If, on the other hand, sequence homology is not found, threading algorithms can be applied. These methods perform pairwise comparison between the target sequence and structural template folds, aiming at discovering if any of the folds can be adopted by the target. The target sequence is therefore threaded through the tertiary structure of protein structure templates. A function measuring the fitness between target and template fold is then calculated and evaluated, with the process being repeated for each template fold. The lowest energy functions are then adopted and modelled in the target structure. In the end, comparative modelling techniques are also applied to access structure integrity [111].

From the overall approaches analyzed, the most successfully implemented methods for structure prediction are homology-based comparative modelling, being Modeller [109], one of the most used and reliable software for performing this task. Table 2.2 summarizes the main available methods and tools for structural prediction in the three discussed methodologies.

Table 2.2 – Depiction of the most important and available tools for protein structural prediction encompassing the three main methodologies, homology modelling, threading and *ab initio* as well as the format in which these tools are available.

Name	Methodology	Format	Reference
Modeller	Homology modelling	Software	[109]
ESyPred3D	Homology modelling	Web-Server	[112]
Swiss-model	Homology modelling	Web-Server	[113]
FoldX	Homology modelling	Software	[114]
WhatIf	Homology modelling	Web-Server	[115]
HHpred	Homology modelling / Threading	Web-Server	[116]
RaptorX	Homology modelling / Threading	Web-Server	[117]
Phyre2	Homology modelling / Threading	Web-Server	[118]
Robetta	Homology modelling/ *ab initio*	Web-Server	[119]
i-tasser	Threading/ *ab initio*	Web-Server	[120]
falcon	Threading/ *ab initio*	Web-Server	[121]
Rosetta@home	*ab initio*	Software	[122]
Evfold	*ab initio*	Web-Server	[123]

Once characterized, either experimentally or modelled, the protein conformation can be observed using protein visualization software. These software showcase the protein structure in a spatial 3D format, often allowing the user to rotate and zoom in specific details of the structure. Several structure representations are also regularly available, from backbone structure to surface display, also allowing secondary structure and van der Waals interaction representation. PyMOL [124], written in the Python programming language, is one of the most used software due to its user-friendly interface and simplicity. This free molecular graphics system not only allows molecular visualization, but also animation and editing. It can also be programmed and extended using Python.

2.3 Machine learning

By making use of the referred methodologies and databases, large amounts of data on enzyme's structural information are possible to be retrieved. This allows for exhaustive analysis of the molecular recognition processes responsible, for example, for the molecular binding of different substrates and cofactors, making way for the elucidation of their specificity mechanisms. However, the analysis of large quantities of protein structures is unmanageable by hand. Due to the inherent characteristics of protein structural data, composed by thousands of atom coordinates for each protein, such analysis requires the input of methodologies for big data analysis. One other hurdle is the fact that important data characteristics and patterns might be missed due to the sheer amount of information available, rendering methodologies capable of learning and interpreting such data an important asset.

Among the most efficient approaches for retrieving seemingly undetectable patterns and characteristics from big data are methods implemented using Machine learning.

Machine learning, a key concept of Artificial Intelligence, is devoted to the development of algorithms that are able to interpret, deduce and generalize new settings of information from data samples, without being explicitly programmed. These algorithms automatically learn and self-improve through experience, resembling the basic concepts of human learning process. They are capable of successfully interpreting unprecedent data by using statistical theory to make inferences from a sample, effectively learning by training.

These methods can also be applied in the automatic mining of large amounts of data in order to extract useful and unknown correlations, creating models able to relate molecular descriptors to biological attributes [125].

The process of machine learning can be divided in two types: supervised and unsupervised learning. For supervised learning, the prediction model is trained using data with known labels, and it can be used to solve classification and regression problems. Classification problems try to identify to which category an object belongs to, while regression tries to predict a continuous-value attribute associated with an object. For unsupervised learning, unlabeled data is used as the input and can be used in clustering, for automatically grouping similar objects into sets [126], [127]. Table 2.3 depicts the referred types of problems and problem characterization along with examples of possible applications and used algorithms.

Table 2.3 – Display of the different type of problems address in machine learning with the corresponding category. Applications and algorithms referent to these problems are also displayed.

Problem	Category	Applications	Algorithms
Classification	Supervised	Classifying two indistinct groups of biological data, Spam detection, Image recognition	Support Vector Machines, nearest neighbors, Naïve bayes, Artificial Neural Network
Regression	Supervised	Drug response, Stock prices	Support vector regression, ridge regression, Lasso
Clustering	Unsupervised	Customer segmentation, Grouping experiment outcomes	k-Means, spectral clustering, mean-shift

Despite the different approaches possible, machine learning algorithms tend to follow the same procedure. First, the data to be analyzed must be prepared for integration and formatted to meet the specificities of the algorithm, being the quality and quantity of the used features an important factor for the success of the method. Afterwards the scoring function of the algorithm is evaluated and optimized. When the model is outputted, statistical and biological interpretations should be performed and, if required, a new optimization iteration should be executed [125].

Support Vector Machines (SVM) are machine learning algorithms based on supervised learning, capable of solving classification problems. These algorithms are among the most used methods in machine learning for addressing biological problems, such us drug discovery [128]–[130] or compound specific activity distinction [131] and accessibility [132].

They work by projecting the integrated data into a high-dimensional space, called hyperspace, and pursuing to find the optimal linear separation state between the features, represented as descriptor vectors, in the hyperplane. As there are an infinite number of possible hyperplane configurations, the algorithm optimizes the separation margin size between classes assuming that larger margins reduce the error of the classifier. Feature points located in both margins are called the "support vectors" [133]. As an example, if given *n* samples, with each having a *m*-dimensional feature vector and one of two classes, such as NAD(H) binding and NADP(H) binding; depending on the quality of the input data, the SVM should be capable of producing a classifier distinguishing different cofactor specificity structures, as depicted in figure 2.3.

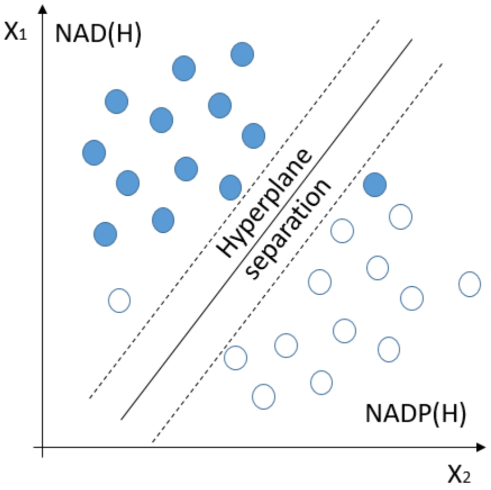


Figure 2.3 – Graphical representation of a SVM. Composing the hyperplane representation are the features (blue for NAD(H) and white for NADP(H), along with the hyperplane separating both groups of features.

With the predictive models generated using the mentioned machine learning algorithms, predicting tools can be developed in order to classify newly collected and unclassified data. The development of such tools can bring great advantages for the fields of systems biology and metabolic engineering as they can be used to guide and predict the outcome of metabolic engineering endeavors and also in the curation process of genome-scale metabolic model (GEM) reconstruction.

In fact, GEM lack the application of such tools in their reconstruction process, as curation steps are performed mainly by literature analysis or experimental characterization. Despite more complete GEMs showing the flow of reducing equivalents and their correlation with the overall metabolism, pointing to possible metabolic engineering targets, the often inaccurate and incorrect cofactor specificity attribution during reconstruction renders the application of these models regularly worthless [39].

2.4 Genome-scale metabolic models

As genomes encode all the information required for the synthesis of enzymes present in cellular metabolism, GEM reconstruction uses this information, along with the attributed metabolic functions of each encoded gene, and corresponds it with the enzymatic reactions catalyzed.

Having become a major tool in systems biology for representing, *in silico*, the set of biochemical events occurring inside an organism's metabolism and respective association of genetic information to these events, GEM reconstruction is nowadays a common practice, with several organisms having their GEM reconstructed. The reconstruction of these models, therefore, is subjected to information on substrates, cofactors and products, as well as stoichiometry and reversibility, from each reaction in the metabolism [134]. Furthermore, information on biomass composition and metabolic energetic requirements are also present.

GEMs are generally used to perform prediction on the phenotype of an organism in response to a set of environmental conditions. Moreover, genomic mutations can also be object of simulation, with phenotypic behavior and flux distribution changes, resulting from gene knockout or overexpression, being analyzed [135], [136]. These capabilities have allowed the usage of these models in the *in silico* optimization and enhancement of several strains for the production of desired compounds [137]. In fact, several metabolic engineering strategies were derived from the simulation of GEMs. A very interesting approach to strain design using GEM was developed by King and Feist [138]. In their work, a method for optimizing *in silico* the cofactor specificity of oxidoreductases was implemented in *Escherichia coli*'s GEM *i*JO1366 [139]. The objective was to modulate cofactor binding specificities in the model in order to enhance bioprocessing strain design predictions. The output was a series of target knockouts and also enzymes for NAD(P)(H) specificity reversal. This type of approach helps evidencing the importance of the machine learning methodologies discussed above for the development of cofactor specificity predicting methods.

GEM reconstruction is nowadays a well-documented process, and several protocols describing the reconstruction process are available [134], [140], [141]. In order to accomplish a correct GEM reconstruction, information on genes, enzyme activity and metabolite transporters must be gathered from genome annotation, being afterwards, the corresponding metabolic and transporting reactions identified and implemented in a reaction network.

Several tool have been developed to aid researchers in this laborious process. Table 2.4 encompass a list of important software used in GEM reconstruction.

Table 2.4 – Display of the most important tools available to perform genome-scale metabolic model reconstruction. A general description of the tool is also depicted.

Software	Description	Reference
merlin	Reconstruction of metabolic networks. Identification of transporter reactions, compartmentalization and organelle localization.	[142]
Biocyc	Repository of Genome Databases containing gene information, transport systems and gap fillers.	[143]
Kbase	Suite of analysis tools and data to support the reconstruction and prediction of metabolic models in microbes and plants.	[144]
Glamm	Reconstruction of metabolic networks from genomic data. Network visualization.	[145]
Raven toolbox	GEM semi-automated reconstruction. Simulation results visualization.	[146]

Tools such as merlin, are an important asset in this process. This user-friendly software performs genome annotations as well as GEM reconstruction. It is able to identify gene metabolic and transporter functions, build the metabolic reaction network, perform model compartmentalization and output a GEM, with the incorporation of a biomass equation [142].

When reconstructed, the model is curated with experimental or literature retrieved data. For validation, GEMs are simulated and their *in silico* biomass growth is compared with experimentally performed growths.

The large majority of developed GEMs use stoichiometric reactions, being their application limited to steady-state modelling of intracellular fluxes [147], [148]. Despite their limitations, these models, also known as constraint-based models, have been applied in many different areas of expertise, such us strain engineering, elucidation of symbiotic relations, pathogenesis and cancer research [148].

Flux balance analysis (FBA) [149] is one of the most used methods for simulating these type of models. This method works by representing the metabolic reactions present in the model as a matrix of stoichiometric coefficients. It then uses linear programming for optimizing the distribution of fluxes in the direction of an objective function, usually biomass growth [150]. This optimization process optimizes the reactions fluxes, which are constrained by upper and lower bounds. In these GEMs the system is assumed as being is steady-state, meaning that the consumption of compounds is equal to their production [149].

Parsimonious FBA (pFBA) [151] is used when a finer refinement of the flux distributions is required. This is achieved by removing futile loops from the network with the minimization of the total sum of fluxes in the network, increasing flux objectivity.

GEMs can be analyzed and simulated using software packages such us OptFlux [152], a very intuitive software, ideal for experimentalist researchers trying to design and develop new applications for their experimental data.

Despite all the efforts and advances made in enhancing GEM reconstruction, as discussed above, several design drawbacks still exist, proving that a much further and precise curation is still in need to achieve a robust and consistent GEM reconstruction capable of reliably predict experimental data using *in silico* methods. Besides the extremely well curated and experimentally characterized GEM reconstruction of *Eschericia coli* [139] and *Saccharomyces cerevisiae* [153], most GEM reconstructions from other organisms contain inconsistencies in their metabolic network composition that impair their application in metabolic engineering and experimental strain design endeavors [154].

2.5 References

[1] J. Watson and F. C. Crick, “MOLECULAR STRUCTURE OF NUCLEIC ACIDS: A Structure for Deoxyribose Nucleic Acid,” *Nature*, no. 4356, p. 4356, 1953.

[2] S. N. Cohen, A. C. Y. Chang, H. W. Boyer, and R. B. Helling, “Construction of Biologically Functional Bacterial Plasmids In Vitro,” *Proc. Natl. Acad. Sci.*, vol. 70, no. 11, pp. 3240–

3244, 1973.

[3] G. Stephanopoulos and J. J. Vallino, "Network Rigidity and Metabolic Engineering in Metabolite Overproduction," *Science (80-.).*, vol. 252, no. 1984, p. 1675, 1991.

[4] J. E. Bailey, "Toward a Science of Metabolic Engineering," *Science (80-.).*, vol. 252, pp. 1668–1697, 1991.

[5] J. W. Lee, D. Na, J. M. Park, J. Lee, S. Choi, and S. Y. Lee, "Systems metabolic engineering of microorganisms for natural and non-natural chemicals," *Nat. Chem. Biol.*, vol. 8, no. 6, pp. 536–546, 2012.

[6] B. M. Woolston, S. Edgar, and G. Stephanopoulos, "Metabolic Engineering: Past and Future," *Annu. Rev. Chem. Biomol. Eng.*, vol. 4, no. 1, pp. 259–288, 2013.

[7] G. Stephanopoulos, "Metabolic Fluxes and Metabolic Engineering," *Metab. Eng.*, vol. 1, no. 1, pp. 1–11, 1999.

[8] V. G. Yadav, M. De Mey, C. Giaw Lim, P. Kumaran Ajikumar, and G. Stephanopoulos, "The future of metabolic engineering and synthetic biology: Towards a systematic practice," *Metab. Eng.*, vol. 14, no. 3, pp. 233–241, 2012.

[9] L. Stein, "Genome annotation: from sequence to biology.," *Nat. Rev. Genet.*, vol. 2, no. July, pp. 493–503, 2001.

[10] H. Kitano, "Systems Biology: A Brief Overview," *Science (80-.).*, vol. 295, no. 5560, pp. 1662–1664, 2002.

[11] D. B. Kell, "Metabolomics and systems biology : making sense of the soup," *Curr. Opin. Microbiol.*, no. 1986, pp. 296–307, 2004.

[12] H. Kitano, "Computational systems biology," *Nature*, vol. 420, no. November, 2002.

[13] B. Ø. Palsson, "Metabolic Systems Biology," *FEBS Lett.*, vol. 583, no. 24, pp. 3900–3904, 2011.

[14] D. Lee, O. Redfern, and C. Orengo, "Predicting protein function from sequence and structure," *Nat. Rev. Mol. Cell Biol.*, vol. 8, no. 12, pp. 995–1005, 2007.

[15] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, "Basic local alignment search tool.," *J. Mol. Biol.*, vol. 215, no. 3, pp. 403–10, 1990.

[16] S. Eddy, "Profile hidden Markov models.," *Bioinformatics*, vol. 14, no. 9, pp. 755–763, 1998.

[17] R. Agarwala, T. Barrett, J. Beck, D. A. Benson, C. Bollin, E. Bolton, D. Bourexis, J. R. Brister, S. H. Bryant, K. Canese, C. Charowhas, K. Clark, M. DiCuccio, I. Dondoshansky, M. Feolo, K. Funk, L. Y. Geer, V. Gorelenkov, W. Hlavina, M. Hoeppner, B. Holmes, M. Johnson, V. Khotomlianski, A. Kimchi, M. Kimelman, P. Kitts, W. Klimke, S. Krasnov, A. Kuznetsov, M. J. Landrum, D. Landsman, J. M. Lee, D. J. Lipman, Z. Lu, T. L. Madden, T. Madej, A. Marchler-Bauer, I. Karsch-Mizrachi, T. Murphy, R. Orris, J. Ostell, C. O'Sullivan, V. Palanigobu, A. R. Panchenko, L. Phan, K. D. Pruitt, K. Rodarmer, W. Rubinstein, E. W. Sayers, V. Schneider, C. L. Schoch, G. D. Schuler, S. T. Sherry, K. Sirotkin, K. Siyan, D. Slotta, A. Soboleva, V. Soussov, G. Starchenko, T. A. Tatusova, K. Todorov, B. W. Trawick, D. Vakatov, Y. Wang, M.

Ward, W. J. Wilbur, E. Yaschenko, and K. Zbicz, “Database Resources of the National Center for Biotechnology Information,” *Nucleic Acids Res.*, vol. 45, no. D1, pp. D12–D17, 2017.

[18] D. Eisenberg, E. M. Marcotte, I. Xenarios, and T. O. Yeates, “Protein function in the post-genomic era.,” *Nature*, vol. 405, no. 6788, pp. 823–6, 2000.

[19] D. Devos and A. Valencia, “Intrinsic errors in genome annotation,” *Trends Genet.*, vol. 17, no. 8, pp. 429–431, 2001.

[20] E. J. Richardson and M. Watson, “The automatic annotation of bacterial genomes,” *Brief. Bioinform.*, vol. 14, no. 1, pp. 1–12, 2013.

[21] P. Stothard and D. S. Wishart, “Automated bacterial genome analysis and annotation,” *Curr. Opin. Microbiol.*, vol. 9, no. 5, pp. 505–510, 2006.

[22] R. Sorek and P. Cossart, “Prokaryotic transcriptomics: A new view on regulation, physiology and pathogenicity,” *Nat. Rev. Genet.*, vol. 11, no. 1, pp. 9–16, 2010.

[23] W. L. Duax, V. Pletnev, A. Addlagatta, J. Bruenn, and C. M. Weeks, “Rational Proteomics I . Fingerprint Identification and Cofactor Specificity in the Short-Chain Oxidoreductase (SCOR) Enzyme Family,” *PROTEINS Struct. Funct. Genet.*, vol. 943, no. February, pp. 931–943, 2003.

[24] J. M. Lee, E. P. Gianchandani, and J. a Papin, “Flux balance analysis in the era of metabolomics.,” *Brief. Bioinform.*, vol. 7, no. 2, pp. 140–50, Jun. 2006.

[25] D. Hilvert, “Design of Protein Catalysts,” *Annu. Rev. Biochem.*, vol. 82, no. 1, pp. 447–470, 2013.

[26] M. Garcia-Viloca, J. Gao, M. Karplus, and D. G. Truhlar, “How Enzymes Work: Analysis by Modern Rate Theory and Computer Simulations,” *Science (80-.).*, vol. 303, no. 5655, pp. 186–195, 2004.

[27] U. T. Bornscheuer, G. W. Huisman, R. J. Kazlauskas, S. Lutz, J. C. Moore, and K. Robins, “Engineering the third wave of biocatalysis,” *Nature*, vol. 485, no. 7397, pp. 185–194, 2012.

[28] S. J. Benkovic and S. Hammes-schiffer, “A Perspective on Enzyme Catalysis,” *Science (80-.).*, vol. 301, no. August, pp. 1196–1202, 2003.

[29] J. L. Faulon, M. Misra, S. Martin, K. Sale, and R. Sapra, “Genome scale enzyme - Metabolite and drug - Target interaction predictions using the signature molecular descriptor,” *Bioinformatics*, vol. 24, no. 2, pp. 225–233, 2008.

[30] W. Tian and J. Skolnick, “How well is enzyme function conserved as a function of pairwise sequence identity?,” *J. Mol. Biol.*, vol. 333, no. 4, pp. 863–882, 2003.

[31] D. Cui, L. Zhang, S. Jiang, Z. Yao, B. Gao, J. Lin, Y. A. Yuan, and D. Wei, “A computational strategy for altering an enzyme in its cofactor preference to NAD(H) and/or NADP(H),” *FEBS J.*, vol. 282, no. 12, pp. 2339–2351, 2015.

[32] A. Bateman, M. J. Martin, C. O’Donovan, M. Magrane, E. Alpi, R. Antunes, B. Bely, M. Bingley, C. Bonilla, R. Britto, B. Bursteinas, H. Bye-AJee, A. Cowley, A. Da Silva, M. De Giorgi, T. Dogan, F. Fazzini, L. G. Castro, L. Figueira, P. Garmiri, G. Georghiou, D. Gonzalez, E. Hatton-

Ellis, W. Li, W. Liu, R. Lopez, J. Luo, Y. Lussi, A. MacDougall, A. Nightingale, B. Palka, K. Pichler, D. Poggioli, S. Pundir, L. Pureza, G. Qi, S. Rosanoff, R. Saidi, T. Sawford, A. Shypitsyna, E. Speretta, E. Turner, N. Tyagi, V. Volynkin, T. Wardell, K. Warner, X. Watkins, R. Zaru, H. Zellner, I. Xenarios, L. Bougueleret, A. Bridge, S. Poux, N. Redaschi, L. Aimo, G. ArgoudPuy, A. Auchincloss, K. Axelsen, P. Bansal, D. Baratin, M. C. Blatter, B. Boeckmann, J. Bolleman, E. Boutet, L. Breuza, C. Casal-Casas, E. De Castro, E. Coudert, B. Cuche, M. Doche, D. Dornevil, S. Duvaud, A. Estreicher, L. Famiglietti, M. Feuermann, E. Gasteiger, S. Gehant, V. Gerritsen, A. Gos, N. Gruaz-Gumowski, U. Hinz, C. Hulo, F. Jungo, G. Keller, V. Lara, P. Lemercier, D. Lieberherr, T. Lombardot, X. Martin, P. Masson, A. Morgat, T. Neto, N. Nouspikel, S. Paesano, I. Pedruzzi, S. Pilbout, M. Pozzato, M. Pruess, C. Rivoire, B. Roechert, M. Schneider, C. Sigrist, K. Sonesson, S. Staehli, A. Stutz, S. Sundaram, M. Tognolli, L. Verbregue, A. L. Veuthey, C. H. Wu, C. N. Arighi, L. Arminski, C. Chen, Y. Chen, J. S. Garavelli, H. Huang, K. Laiho, P. McGarvey, D. A. Natale, K. Ross, C. R. Vinayaka, Q. Wang, Y. Wang, L. S. Yeh, and J. Zhang, "UniProt: The universal protein knowledgebase," *Nucleic Acids Res.*, vol. 45, no. D1, pp. D158–D169, 2017.

[33] S. Placzek, I. Schomburg, A. Chang, L. Jeske, M. Ulbrich, J. Tillack, and D. Schomburg, "BRENDA in 2017: New perspectives and new tools in BRENDA," *Nucleic Acids Res.*, vol. 45, no. D1, pp. D380–D388, 2017.

[34] M. Kanehisa, M. Furumichi, M. Tanabe, Y. Sato, and K. Morishima, "KEGG: New perspectives on genomes, pathways, diseases and drugs," *Nucleic Acids Res.*, vol. 45, no. D1, pp. D353–D361, 2017.

[35] I. V. Tetko, I. V. Rodchenkov, M. C. Walter, T. Rattei, and H. W. Mewes, "Beyond the 'best' match: Machine learning annotation of protein sequences by integration of different sources of information," *Bioinformatics*, vol. 24, no. 5, pp. 621–628, 2008.

[36] V. Volpato, A. Adelfio, and G. Pollastri, "Accurate prediction of protein enzymatic class by N-to-1 Neural Networks," *BMC Bioinformatics*, vol. 14, no. Suppl 1, pp. 1–7, 2013.

[37] Y. Gao, B. Li, Y. Cai, K. Feng, Z. Li, and Y. Jiang, "Prediction of active sites of enzymes by maximum relevance minimum redundancy (mRMR) feature selection," *Mol. Biosyst.*, vol. 9, pp. 61–69, 2013.

[38] Y. Matsuta, M. Ito, and Y. Tohsato, "ECOH : An Enzyme Commission number predictor using mutual information and a support vector machine," *Bioinformatics*, vol. 29, no. 3, pp. 365–372, 2013.

[39] Y. Wang, K. Y. San, and G. N. Bennett, "Cofactor engineering for advancing chemical biotechnology," *Curr. Opin. Biotechnol.*, vol. 24, no. 6, pp. 994–999, 2013.

[40] H. Zhao and W. A. Van Der Donk, "Regeneration of cofactors for use in biocatalysis," *Curr. Opin. Biotechnol.*, vol. 14, no. 6, pp. 583–589, 2003.

[41] J. C. Xavier, K. R. Patil, and I. Rocha, "Integration of Biomass Formulations of Genome-Scale Metabolic Models with Experimental Data Reveals Universally Essential Cofactors in Prokaryotes," *Metab. Eng.*, vol. 39, no. October 2016, pp. 200–208, 2017.

[42] W. Liu and P. Wang, "Cofactor regeneration for sustainable enzymatic biosynthesis," *Biotechnol. Adv.*, vol. 25, no. 4, pp. 369–384, 2007.

[43] C. R. Shen, E. I. Lan, Y. Dekishima, A. Baez, K. M. Cho, and J. C. Liao, “Driving forces enable high-titer anaerobic 1-butanol synthesis in Escherichia coli,” *Appl. Environ. Microbiol.*, vol. 77, no. 9, pp. 2905–2915, 2011.

[44] G. C. Zhang, J. J. Liu, and W. T. Ding, “Decreased xylitol formation during xylose fermentation in saccharomyces cerevisiae due to overexpression of water-forming NADH oxidase,” *Appl. Environ. Microbiol.*, vol. 78, no. 4, pp. 1081–1086, 2012.

[45] J. W. Kim, Y. W. Chin, Y. C. Park, and J. H. Seo, “Effects of deletion of glycerol-3-phosphate dehydrogenase and glutamate dehydrogenase genes on glycerol and ethanol metabolism in recombinant Saccharomyces cerevisiae,” *Bioprocess Biosyst. Eng.*, vol. 35, no. 1–2, pp. 49–54, 2012.

[46] H. TAMAKAWA, S. IKUSHIMA, and S. YOSHIDA, “Ethanol Production from Xylose by a Recombinant *Candida utilis* Strain Expressing Protein-Engineered Xylose Reductase and Xylitol Dehydrogenase,” *Biosci. Biotechnol. Biochem.*, vol. 75, no. 10, pp. 1994–2000, 2011.

[47] A. K. Bera, N. W. Y. Ho, A. Khan, and M. Sedlak, “A genetic overhaul of Saccharomyces cerevisiae 424A(LNH-ST) to improve xylose fermentation,” *J. Ind. Microbiol. Biotechnol.*, vol. 38, no. 5, pp. 617–626, 2011.

[48] Z. P. Guo, L. Zhang, Z. Y. Ding, Z. X. Wang, and G. Y. Shi, “Improving ethanol productivity by modification of glycolytic redox factor generation in glycerol-3-phosphate dehydrogenase mutants of an industrial ethanol yeast,” *J. Ind. Microbiol. Biotechnol.*, vol. 38, no. 8, pp. 935–943, 2011.

[49] Z. peng Guo, L. Zhang, Z. yang Ding, and G. yang Shi, “Minimization of glycerol synthesis in industrial ethanol yeast without influencing its fermentation performance,” *Metab. Eng.*, vol. 13, no. 1, pp. 49–59, 2011.

[50] R. Hector, J. Mertens, M. Bowman, N. Nichols, and M. Cotta, “Saccharomyces cerevisiae engineered for xylose metabolism requires gluconeogenesis and the oxidative branch of the pentose phosphate pathway for aerobic xylose assimilation,” *Yeast*, vol. 28, no. 10, pp. 645–660, 2011.

[51] S. Q. Pham, P. Gao, and Z. Li, “Engineering of recombinant E. coli cells co-expressing P450pyrTM monooxygenase and glucose dehydrogenase for highly regio- and stereoselective hydroxylation of alicycles with cofactor recycling,” *Biotechnol. Bioeng.*, vol. 110, no. 2, pp. 363–373, 2013.

[52] Y. G. Kim, S. Lee, O. S. Kwon, S. Y. Park, S. J. Lee, B. J. Park, and K. J. Kim, “Redox-switch modulation of human SSADH by dynamic catalytic loop,” *EMBO J.*, vol. 28, no. 7, pp. 959–968, 2009.

[53] S. G. Bell, F. Xu, and E. O. D. Johnson, “Protein recognition in ferredoxin – P450 electron transfer in the class I CYP199A2 system from Rhodopseudomonas palustris,” *J Biol Inorg Chem*, vol. 15, pp. 315–328, 2010.

[54] Z. Xiao, C. Lv, C. Gao, J. Qin, C. Ma, Z. Liu, P. Liu, L. Li, and P. Xu, “A novel whole-cell biocatalyst with NAD+ regeneration for production of chiral chemicals,” *PLoS One*, vol. 5, no.

1, pp. 1–6, 2010.

[55] L. Wang, H. Zhang, C. B. Ching, Y. Chen, and R. Jiang, “Nanotube-supported bioproduction of 4-hydroxy-2-butanone via in situ cofactor regeneration,” *Appl. Microbiol. Biotechnol.*, vol. 94, no. 5, pp. 1233–1241, 2012.

[56] A. B. Canelas, W. M. Van Gulik, and J. J. Heijnen, “Determination of the cytosolic free NAD/NADH ratio in Saccharomyces cerevisiae under steady-state and highly dynamic conditions,” *Biotechnol. Bioeng.*, vol. 100, no. 4, pp. 734–743, 2008.

[57] J. P. van Dijken and W. A. Scheffers, “Redox balances in the metabolism of sugars by yeasts,” *FEMS Microbiol. Lett.*, vol. 32, no. 3–4, pp. 199–224, 1986.

[58] J. Pain, M. M. Balamurali, A. Dancis, and D. Pain, “Mitochondrial NADH kinase, Pos5p, is required for efficient iron-sulfur cluster biogenesis in Saccharomyces cerevisiae,” *J. Biol. Chem.*, vol. 285, no. 50, pp. 39409–39424, 2010.

[59] F. Shi, Z. Li, M. Sun, and Y. Li, “Role of mitochondrial NADH kinase and NADPH supply in the respiratory chain activity of Saccharomyces cerevisiae,” *Acta Biochim. Biophys. Sin. (Shanghai).*, vol. 43, no. 12, pp. 989–995, 2011.

[60] J. Hou, G. N. Vemuri, X. Bao, and L. Olsson, “Impact of overexpressing NADH kinase on glucose and xylose metabolism in recombinant xylose-utilizing Saccharomyces cerevisiae,” *Appl. Microbiol. Biotechnol.*, vol. 82, no. 5, pp. 909–919, 2009.

[61] Y. H. Hua, C. Y. Wu, K. Sargsyan, and C. Lim, “Sequence-motif Detection of NAD(P)-binding Proteins: Discovery of a Unique Antibacterial Drug Target,” *Sci. Rep.*, vol. 4, pp. 1–7, 2014.

[62] O. Carugo and P. Argos, “NADP-dependent enzymes. I: Conserved stereochemistry of cofactor binding,” *Proteins Struct. Funct. Genet.*, vol. 28, no. 1, pp. 10–20, 1997.

[63] H. M. Geertz-Hansen, N. Blom, A. M. Feist, S. Brunak, and T. N. Petersen, “Cofactory: Sequence-based prediction of cofactor specificity of Rossmann folds,” *Proteins Struct. Funct. Bioinforma.*, vol. 82, no. 9, pp. 1819–1828, 2014.

[64] J. K. B. Cahn, S. Brinkmann-Chen, T. Spatzal, J. A. Wiig, A. R. Buller, O. Einsle, Y. Hu, M. W. Ribbe, and F. H. Arnold, “Cofactor specificity motifs and the induced fit mechanism in class I ketol-acid reductoisomerases,” *Biochem. J.*, vol. 468, no. 3, pp. 475–484, 2015.

[65] L. S. Vidal, C. L. Kelly, P. M. Mordaka, and J. T. Heap, “Review of NAD (P) H-dependent oxidoreductases : Properties , engineering and application,” *BBA - Proteins Proteomics*, vol. 1866, no. August 2017, pp. 327–347, 2018.

[66] S. T. Rao and M. G. Rossmann, “Comparison of super-secondary structures in proteins,” *J. Mol. Biol.*, vol. 76, no. 2, pp. 241–256, 1973.

[67] C. I. Brändén, “The TIM barrel-the most frequently occurring folding motif in proteins. Current Opinion in Structural Biology 1991, 1:978-983,” *Curr. Opin. Struct. Biol.*, vol. 1, no. 6, pp. 978–983, 1991.

[68] J. K. B. Cahn, S. Brinkmann-Chen, A. R. Buller, and F. H. Arnold, “Artificial domain duplication replicates evolutionary history of ketol-acid reductoisomerases,” *Protein Sci.*, vol.

25, pp. 1241–1248, 2016.

[69] G. Kleiger and D. Eisenberg, "GXXXG and GXXXA motifs stabilize FAD and NAD(P)-binding rossmann folds through Cα-H···O hydrogen bonds and van der Waals interactions," *J. Mol. Biol.*, vol. 323, no. 1, pp. 69–76, 2002.

[70] C. Y. Wu, Y. H. Hwa, Y. C. Chen, and C. Lim, "Hidden relationship between conserved residues and locally conserved phosphate-binding structures in NAD(P)-binding proteins," *J. Phys. Chem. B*, vol. 116, no. 19, pp. 5644–5652, 2012.

[71] C. C. Marohnic, M. C. Bewley, and M. J. Barber, "Engineering and characterization of a NADPH-utilizing cytochrome b 5 reductase," *Biochemistry*, vol. 42, no. 38, pp. 11170–11182, 2003.

[72] G. A. Khoury, H. Fazelinia, J. W. Chin, R. J. Pantazes, P. C. Cirino, and C. D. Maranas, "Computational design of Candida boidinii xylose reductase for altered cofactor specificity," *Protein Sci.*, vol. 18, no. 10, pp. 2125–2138, 2009.

[73] M. Nishiyama, J. J. Birktoft, and T. Beppu, "Alteration of coenzyme specificity of malate dehydrogenase from Thermus flavus by site-directed mutagenesis," *J. Biol. Chem.*, vol. 268, no. 7, pp. 4656–4660, 1993.

[74] N. Holmberg, U. Ryde, and L. Bülow, "Redesign of the coenzyme specificity in l-Lactate dehydrogenase from Bacillus stearothermophilus using site-directed mutagenesis and media engineering," *Protein Eng. Des. Sel.*, vol. 12, no. 10, pp. 851–856, 1999.

[75] S. P. Miller, M. Lunzer, and A. M. Dean, "Direct demonstration of an adaptive constraint," *Science (80-.).*, vol. 314, no. 5798, pp. 458–461, 2006.

[76] J. a Bocanegra, N. S. Scrutton, and R. N. Perham, "Creation of an NADP-dependent pyruvate dehydrogenase multienzyme complex by protein engineering.," *Biochemistry*, vol. 32, no. 11, pp. 2737–2740, 1993.

[77] N. Bernard, K. Johnsen, J. Holbrook, and J. Delcour, "D175 discriminates between NADH and NADPH in the coenzyme binding site of Lactobacillus delbrueckii subsp. bulgaricus D-lactate dehydrogenase," *Biochem. Biophys. Res. Commun.*, vol. 208, no. 3, pp. 895–900, 1995.

[78] S. Clermont, C. Corbier, Y. Mely, D. Gerard, A. Wonacott, and G. Branlant, "Determinants of Coenzyme Specificity in Glyceraldehyde-3-phosphate Dehydrogenase: Role of the Acidic Residue in the Fingerprint Region of the Nucleotide Binding Fold," *Biochemistry*, vol. 32, no. 38, pp. 10178–10184, 1993.

[79] A. H. Ehrensberger, R. A. Elling, and D. K. Wilson, "Structure-guided engineering of xylitol dehydrogenase cosubstrate specificity," *Structure*, vol. 14, no. 3, pp. 567–575, 2006.

[80] J. Y. Hsieh, G. Y. Liu, G. G. Chang, and H. C. Hung, "Determinants of the dual cofactor specificity and substrate cooperativity of the human mitochondrial NAD(P)+-dependent malic enzyme: Functional roles of glutamine 362," *J. Biol. Chem.*, vol. 281, no. 32, pp. 23237–23245, 2006.

[81] S. Watanabe, T. Kodaki, and K. Makino, "Complete reversal of coenzyme specificity of xylitol

dehydrogenase and increase of thermostability by the introduction of structural zinc," *J. Biol. Chem.*, vol. 280, no. 11, pp. 10340–10349, 2005.

[82] R. Woodyer, W. A. Van der Donk, and H. Zhao, "Relaxing the nicotinamide cofactor specificity of phosphite dehydrogenase by rational design," *Biochemistry*, vol. 42, no. 40, pp. 11604–11614, 2003.

[83] A. Serov, A. Popova, V. Fedorchuk, and V. Tishkov, "Engineering of coenzyme specificity of formate dehydrogenase from Saccharomyces cerevisiae Alexander," *Biochem. J.*, vol. 367, pp. 841–847, 2002.

[84] R. Chen, A. F. Greer, A. M. Dean, and J. H. Hurley, "Redesigning secondary structure to invert coenzyme specificity in isopropylmalate dehydrogenase," *Proc. Natl. Acad. Sci.*, vol. 93, pp. 12171–12176, 1996.

[85] a Galkin, L. Kulakova, T. Ohshima, N. Esaki, and K. Soda, "Construction of a new leucine dehydrogenase with preferred specificity for NADP+ by site-directed mutagenesis of the strictly NAD+-specific enzyme.," *Protein Eng.*, vol. 10, no. 6, pp. 687–690, 1997.

[86] J. A. Friesen, C. Martin Lawrence, C. V. Stauffacher, and V. W. Rodwell, "Structural determinants of nucleotide coenzyme specificity in the distinctive dinucleotide binding fold of HMG-CoA reductase from Pseudomonas mevalonii," *Biochemistry*, vol. 35, no. 37, pp. 11945–11950, 1996.

[87] S. Leitgeb, B. Petschacher, D. K. Wilson, and B. Nidetzky, "Fine tuning of coenzyme specificity in family 2 aldo-keto reductases revealed by crystal structures of the Lys-274 ??? Arg mutant of Candida tenuis xylose reductase (AKR2B5) bound to NAD+ and NADP+," *FEBS Lett.*, vol. 579, no. 3, pp. 763–767, 2005.

[88] B. PETSCHACHER, S. LEITGEB, K. L. KAVANAGH, D. K. WILSON, and B. NIDETZKY, "The coenzyme specificity of Candida tenuis xylose reductase (AKR2B5) explored by site-directed mutagenesis and X-ray crystallography," *Biochem. J.*, vol. 385, no. 1, pp. 75–83, 2005.

[89] S. Banta, B. A. Swanson, S. Wu, A. Jarnagin, and S. Anderson, "Optimizing an artificial metabolic pathway: Engineering the cofactor specificity of Corynebacterium 2,5-diketo-D-gluconic acid reductase for use in vitamin C biosynthesis," *Biochemistry*, vol. 41, no. 20, pp. 6226–6236, 2002.

[90] S. Banta, B. A. Swanson, S. Wu, A. Jarnagin, and S. Anderson, "Alteration of the specificity of the cofactor-binding pocket of Corynebacterium 2,5-diketo-D-gluconic acid reductase A," *Protein Eng.*, vol. 15, no. 2, pp. 131–140, 2002.

[91] N. S. Scrutton, A. Berry, and R. N. Perham, "Redesign of the coenzyme specificity of a dehydrogenase by protein engineering," *Nature*, vol. 343, no. 6253, pp. 38–43, 1990.

[92] M. J. Rane and K. C. Calvo, "Reversal of the nucleotide specificity of ketol acid reductoisomerase by site-directed mutagenesis identifies the NADPH binding site," *Arch. Biochem. Biophys.*, vol. 338, no. 1, pp. 83–89, 1997.

[93] N. Shiraishi, C. Croy, J. Kaur, and W. H. Campbell, "Engineering of Pyridine Nucleotide Specificity of Nitrate Reductase: Mutagenesis of Recombinant CytochromebReductase Fragment ofNeurospora crassaNADPH:Nitrate Reductase," *Arch. Biochem. Biophys.*, vol.

358, no. 1, pp. 104–115, 1998.

[94] M. Kostrzynska, C. R. Sopher, and H. Lee, “Mutational analysis of the role of the conserved lysine-270 in the _Pichia stipitis_ xylose reductase.,” *FEMS Microbiol.Lett.*, vol. 159, no. March, pp. 107–112, 1998.

[95] L. Liang, J. Zhang, and Z. Lin, “Altering coenzyme specificity of Pichia stipitis xylose reductase by the semi-rational approach CASTing,” *Microb. Cell Fact.*, vol. 6, pp. 1–11, 2007.

[96] M. H. M. Eppink, K. M. Overkamp, H. A. Schreuder, and W. J. H. Van Berkel, “Switch of coenzyme specificity of p-hydroxybenzoate hydroxylase,” *J. Mol. Biol.*, vol. 292, no. 1, pp. 87–96, 1999.

[97] C. L. Elmore and T. D. Porter, “Modification of the nucleotide cofactor-binding site of cytochrome P-450 reductase to enhance turnover with NADH in vivo,” *J. Biol. Chem.*, vol. 277, no. 50, pp. 48960–48964, 2002.

[98] K. Kristan, J. Stojan, J. Adamski, and T. Lanišnik Rižner, “Rational design of novel mutants of fungal 17β-hydroxysteroid dehydrogenase,” *J. Biotechnol.*, vol. 129, no. 1, pp. 123–130, 2007.

[99] T. R. Dambe, A. M. Kühn, T. Brossette, F. Giffhorn, and A. J. Scheidig, “Crystal structure of NADP(H)-dependent 1,5-anhydro-D-fructose reductase from Sinorhizobium morelense at 2.2 Å resolution: Construction of a NADH-accepting mutant and its application in rare sugar synthesis,” *Biochemistry*, vol. 45, no. 33, pp. 10030–10042, 2006.

[100] M. Medina, A. Luquita, J. Tejero, J. Hermoso, T. Mayoral, J. Sanz-Aparicio, K. Grever, and C. Gómez-Moreno, “Probing the Determinants of Coenzyme Specificity in Ferredoxin-NADP+Reductase by Site-directed Mutagenesis,” *J. Biol. Chem.*, vol. 276, no. 15, pp. 11902–11912, 2001.

[101] R. Chen, a Greer, and a M. Dean, “A highly active decarboxylating dehydrogenase with rationally inverted coenzyme specificity.,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 92, no. 25, pp. 11666–11670, 1995.

[102] T. Yaoi, K. Miyazaki, T. Oshima, Y. Komukai, and M. Go, “Conversion of the coenzyme specificity of isocitrate dehydrogenase by module replacement,” *J. Biochem.*, vol. 119, no. 5, pp. 1014–1018, 1996.

[103] L. Zhang, B. Ahvazi, R. Szittner, A. Vrielink, and E. Meighen, “Change of nucleotide specificity and enhancement of catalytic efficiency in single point mutants of Vibrio harveyi aldehyde dehydrogenase,” *Biochemistry*, vol. 38, no. 35, pp. 11440–11447, 1999.

[104] A. Rodríguez-Arnedo, M. Camacho, F. Llorca, and M. J. Bonete, “Complete reversal of coenzyme specificity of isocitrate dehydrogenase from Haloferax volcanii,” *Protein J.*, vol. 24, no. 5, pp. 259–266, 2005.

[105] C. A. F. George A. Khoury, James Smadbeck, Chris A. Kieslich, “Protein folding and de novo protein design for biotechnological applications,” *Trends Biotechnol*, vol. 32, no. 2, pp. 99–109, 2015.

[106] C. Levinthal, “How to fold graciously,” *Mössbauer Spectrosc. Biol. Syst. Proc.*, vol. 24, no.

41, pp. 22–24, 1969.

[107] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, "The protein data bank.," *Nucleic Acids Res.*, vol. 28, no. 1, pp. 235–242, 2000.

[108] R. Bonneau and D. Baker, "Ab inition protein structure prediction: progress and prospects," *Annu. Rev. Biophys. Biomol. Struct.*, vol. 30, pp. 173–189, 2001.

[109] B. Webb and A. Sali, *Comparative protein structure modeling using MODELLER*, vol. 2014. 2014.

[110] M. A. Marti-Renom, A. C. Stuart, R. Sanchez, F. Melo, and A. Sali, "Comparative Protein Structure Modeling of Genes and Genomes," *Annu. Rev. Biophys. Biomol. Struct.*, vol. 29, pp. 291–325, 2000.

[111] J. Schonbrun, W. J. Wedemeyer, and D. Baker, "Protein structure prediction in 2002," *Curr. Opin. Struct. Biol.*, vol. 12, no. 3, pp. 348–354, 2002.

[112] C. Lambert, N. Léonard, X. De Bolle, and E. Depiereux, "ESyPred3D: Prediction of proteins 3D structures," *Bioinformatics*, vol. 18, no. 9, pp. 1250–1256, 2002.

[113] S. Bienert, A. Waterhouse, T. A. P. De Beer, G. Tauriello, G. Studer, L. Bordoli, and T. Schwede, "The SWISS-MODEL Repository-new features and functionality," *Nucleic Acids Res.*, vol. 45, no. D1, pp. D313–D319, 2017.

[114] P. Vanhee, E. Verschueren, L. Baeten, F. Stricher, L. Serrano, F. Rousseau, and J. Schymkowitz, "BriX: A database of protein building blocks for structural analysis, modeling and design," *Nucleic Acids Res.*, vol. 39, no. SUPPL. 1, pp. 435–442, 2011.

[115] G. Vriend, "WHAT IF: A molecular modeling and drug design program," *J. Mol. Graph.*, vol. 8, no. 1, pp. 52–56, 1990.

[116] J. Söding, A. Biegert, and A. N. Lupas, "The HHpred interactive server for protein homology detection and structure prediction," *Nucleic Acids Res.*, vol. 33, no. SUPPL. 2, pp. 244–248, 2005.

[117] S. Wang, W. Li, S. Liu, and J. Xu, "RaptorX-Property: a web server for protein structure property prediction," *Nucleic Acids Res.*, vol. 44, no. W1, pp. W430–W435, 2016.

[118] L. A. Kelly, S. Mezulis, C. Yates, M. Wass, and M. Sternberg, "The Phyre2 web portal for protein modelling, prediction, and analysis," *Nat. Protoc.*, vol. 10, no. 6, pp. 845–858, 2015.

[119] D. E. Kim, D. Chivian, and D. Baker, "Protein structure prediction and analysis using the Robetta server," *Nucleic Acids Res.*, vol. 32, no. WEB SERVER ISS., pp. 526–531, 2004.

[120] Y. Zhang, "I-TASSER server for protein 3D structure prediction," *BMC Bioinformatics*, vol. 9, pp. 1–8, 2008.

[121] S. C. Li, D. Bu, J. Xu, and M. Li, "Fragment-HMM: A new approach to protein structure prediction," *Protein Sci.*, vol. 17, p. 1925, 2008.

[122] R. Das and D. Baker, "Macromolecular Modeling with Rosetta," *Annu. Rev. Biochem.*, vol.

77, no. 1, pp. 363–382, 2008.

[123] R. Sheridan, R. J. Fieldhouse, S. Hayat, Y. Sun, Y. Antipin, L. Yang, T. Hopf, D. S. Marks, and C. Sander, "EVfold.org: Evolutionary Couplings and Protein 3D Structure Prediction," pp. 1–17, 2015.

[124] W. DeLano, "Pymol: An open-source molecular graphics tool," *CCP4 Newsl. Protein Crystallogr.*, vol. 700, 2002.

[125] P. Larrañaga, B. Calvo, R. Santana, C. Bielza, J. Galdiano, I. Inza, J. A. Lozano, R. Armañanzas, G. Santafé, A. Pérez, and V. Robles, "Machine learning in bioinformatics," *Brief. Bioinform.*, vol. 7, no. 1, pp. 86–112, 2006.

[126] H. Ding, I. Takigawa, H. Mamitsuka, and S. Zhu, "Similarity-based machine learning methods for predicting drug-target interactions: A brief review," *Brief. Bioinform.*, vol. 15, no. 5, pp. 734–747, 2013.

[127] A. Lavecchia, "Machine-learning approaches in drug discovery: Methods and applications," *Drug Discov. Today*, vol. 20, no. 3, pp. 318–331, 2015.

[128] E. Byvatov, U. Fechner, J. Sadowski, and G. Schneider, "Comparison of Support Vector Machine and Artificial Neural Network Systems for Drug/Nondrug Classification," *J. Chem. Inf. Model.*, vol. 43, no. 6, pp. 1882–1889, 2003.

[129] V. V. Zernov, K. V. Balakin, A. A. Ivaschenko, N. P. Savchuk, and I. V. Pletnev, "Drug Discovery Using Support Vector Machines. The Case Studies of Drug-likeness, Agrochemical-likeness, and Enzyme Inhibition Predictions," *J. Chem. Inf. Comput. Sci.*, vol. 43, no. 6, pp. 2048–2056, 2003.

[130] M. K. Warmuth, J. Liao, G. Ratsch, M. Mathieson, S. Putta, and C. Lemmen, "Active learning with support vector machines in the drug discovery process," *J Chem Inf Comput Sci*, vol. 43, pp. 667–673, 2003.

[131] R. N. Jorissen and M. K. Gilson, "Virtual Screening of Molecular Databases Using a Support Vector Machine," *Society*, vol. 45, pp. 549–561, 2005.

[132] Y. Podolyan, M. a Walters, and G. Karypis, "Assessing synthetic accessibility of chemical compounds using machine learning methods.," *J Chem Inf Model*, vol. 50, no. 6, pp. 979–991, 2010.

[133] C. Cortes and V. Vapnik, "Support Vector Networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, 1995.

[134] A. M. Feist, M. J. Herrgård, I. Thiele, J. L. Reed, and B. Palsson, "Reconstruction of biochemical networks in microorganisms," *Nat. Rev. Microbiol.*, vol. 7, no. 2, pp. 129–143, 2009.

[135] K. R. Patil, M. Åkesson, and J. Nielsen, "Use of genome-scale microbial models for metabolic engineering," *Curr. Opin. Biotechnol.*, vol. 15, no. Figure 1, pp. 64–69, 2004.

[136] C. B. Milne, P. J. Kim, J. A. Eddy, and N. D. Price, "Accomplishments in genome-scale in silico modeling for industrial and medical biotechnology," *Biotechnol. J.*, vol. 4, no. 12, pp.

1653–1670, 2009.

[137] A. P. Burgard, P. Pharkya, and C. D. Maranas, "OptKnock: A Bilevel Programming Framework for Identifying Gene Knockout Strategies for Microbial Strain Optimization," *Biotechnol. Bioeng.*, vol. 84, no. 6, pp. 647–657, 2003.

[138] Z. A. King and A. M. Feist, "Optimizing Cofactor Specificity of Oxidoreductase Enzymes for the Generation of Microbial Production Strains—OptSwap," *Ind. Biotechnol.*, vol. 9, no. 4, pp. 236–246, 2013.

[139] J. D. Orth, T. M. Conrad, J. Na, J. A. Lerman, H. Nam, A. M. Feist, and B. Palsson, "A comprehensive genome-scale reconstruction of Escherichia coli metabolism-2011," *Mol. Syst. Biol.*, vol. 7, no. 535, pp. 1–9, 2011.

[140] I. Rocha, J. Förster, and J. Nielsen, "Design and application of genome-scale reconstructed metabolic models," in *Methods in molecular biology (Clifton, N.J.)*, vol. 416, 2007, pp. 409–433.

[141] I. Thiele and B. Palsson, "A protocol for generating a high-quality genome-scale metabolic reconstruction," *Nat Protoc*, vol. 5, no. 1, pp. 93–121, 2010.

[142] O. Dias, R. Pereira, A. K. Gombert, E. C. Ferreira, and I. Rocha, "iOD907, the first genome-scale metabolic model for the milk yeast Kluyveromyces lactis," *Biotechnol. J.*, vol. 9, pp. 776–790, 2014.

[143] R. Caspi, T. Altman, R. Billington, K. Dreher, H. Foerster, C. a. Fulcher, T. a. Holland, I. M. Keseler, A. Kothari, A. Kubo, M. Krummenacker, M. Latendresse, L. a. Mueller, Q. Ong, S. Paley, P. Subhraveti, D. S. Weaver, D. Weerasinghe, P. Zhang, and P. D. Karp, "The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases," *Nucleic Acids Res.*, vol. 42, no. October 2007, pp. 623–631, 2014.

[144] R. Overbeek, T. Begley, R. M. Butler, J. V. Choudhuri, H. Y. Chuang, M. Cohoon, V. de Crécy-Lagard, N. Diaz, T. Disz, R. Edwards, M. Fonstein, E. D. Frank, S. Gerdes, E. M. Glass, A. Goesmann, A. Hanson, D. Iwata-Reuyl, R. Jensen, N. Jamshidi, L. Krause, M. Kubal, N. Larsen, B. Linke, A. C. McHardy, F. Meyer, H. Neuweger, G. Olsen, R. Olson, A. Osterman, V. Portnoy, G. D. Pusch, D. a. Rodionov, C. Rül;ckert, J. Steiner, R. Stevens, I. Thiele, O. Vassieva, Y. Ye, O. Zagnitko, and V. Vonstein, "The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes," *Nucleic Acids Res.*, vol. 33, no. 17, pp. 5691–5702, 2005.

[145] J. T. Bates, D. Chivian, and A. P. Arkin, "GLAMM: Genome-Linked Application for Metabolic Maps," *Nucleic Acids Res.*, vol. 39, no. SUPPL. 2, pp. 400–405, 2011.

[146] R. Agren, L. Liu, S. Shoaie, W. Vongsangnak, I. Nookaew, and J. Nielsen, "The RAVEN Toolbox and Its Use for Generating a Genome-scale Metabolic Model for Penicillium chrysogenum," *PLoS Comput. Biol.*, vol. 9, no. 3, 2013.

[147] A. K. Gombert and J. Nielsen, "Mathematical modelling of metabolism," *Curr. Opin. Biotechnol.*, vol. 11, no. 2, pp. 180–186, 2000.

[148] A. Bordbar, J. M. Monk, Z. A. King, and B. O. Palsson, "Constraint-based models predict metabolic and associated cellular functions," *Nat. Rev. Genet.*, vol. 15, no. 2, pp. 107–120,

2014.

[149] J. D. Orth, I. Thiele, and B. Ø. Palsson, “What is flux balance analysis?,” *Nat Biotechnol*, vol. 28, no. 3, pp. 245–248, 2010.

[150] J. S. Edwards and B. O. Palsson, “Systems Properties of the Haemophilus influenzae Rd Metabolic Genotype Systems Properties of the Haemophilus influenzae Rd Metabolic Genotype,” *cell Biol. Metab.*, vol. 274, pp. 17410–17416, 1999.

[151] N. E. Lewis, K. K. Hixson, T. M. Conrad, J. A. Lerman, P. Charusanti, A. D. Polpitiya, J. N. Adkins, G. Schramm, S. O. Purvine, D. Lopez-Ferrer, K. K. Weitz, R. Eils, R. König, R. D. Smith, and B. Palsson, “Omic data from evolved E. coli are consistent with computed optimal growth from genome-scale models,” *Mol. Syst. Biol.*, vol. 6, no. 390, 2010.

[152] I. Rocha, P. Maia, P. Evangelista, P. Vilaça, S. Soares, J. P. Pinto, J. Nielsen, K. R. Patil, E. C. Ferreira, and M. Rocha, “OptFlux: an open-source software platform for in silico metabolic engineering.,” *BMC Syst. Biol.*, vol. 4, p. 45, 2010.

[153] H. W. Aung, S. A. Henry, and L. P. Walker, “Revising the Representation of Fatty Acid, Glycerolipid, and Glycerophospholipid Metabolism in the Consensus Model of Yeast Metabolism,” *Ind. Biotechnol.*, vol. 9, no. 4, pp. 215–228, 2013.

[154] C. Zhang and Q. Hua, “Applications of genome-scale metabolic models in biotechnology and systems medicine,” *Front. Physiol.*, vol. 6, no. JAN, pp. 1–8, 2016.

Chapter 3

Unveiling The Molecular Determinants For NAD(P)(H) Cofactor Specificity Using Structural Information

The information presented in this Chapter is being prepared for submission to a peer reviewed journal:

Resende T., Soares C., Rocha I. Unveiling the molecular determinants for NAD(P)(H) cofactor specificity using structural information.

3.1 Introduction

Systems biology foundations broadly rely upon well performed gene annotations [1]. In this post-genomic era, given the large and exponentially growing amount of sequences being characterized, experimental determination of a protein's function is becoming unfeasible, due to its cost and time consumption [2]–[4]. Current methodologies are based on pairwise sequence alignment and search for sequence homology to perform protein function annotation [5], [6]. However, the usage of such approaches in annotation pipelines tend to continuously propagate annotation errors across all sequenced organisms due to the attribution of outdated and unspecific functions to new annotated genes, impairing the discovery of new gene functions [7]. Despite their usefulness and relevance, such methodologies fail in capturing essential information hidden in dissimilar areas of different sequences, such as cofactor specificity, which gravely impairs the understanding of an organism's metabolism.

Notwithstanding the utterly massive increment of protein sequences available in public databases, only a residual amount have information on their biological functions reviewed. The Uniprot database [8] release from October 25th, 2017, has indexed approximately 93 million protein sequences, with less than 0.6% manually annotated and reviewed, being the vast majority automatically annotated. Brenda-enzymes [9], an on-line database containing curated experimental information on enzymes and enzyme activity, only encompassed data on a total of 7270 different enzymes in October 30th, 2017. This vast gap between protein sequence and biological processes prejudices the assignment of protein function, with particular effects on the accurate identification of cofactor and substrate usage in metabolic reactions.

Cofactors act in enzymatic reactions as redox carries and are important mediators for energy transfer in the cell. The lack of accuracy in determining cofactor usage in numerous genes severely affects, for example, genome-scale metabolic model reconstruction, as well as metabolic engineering and strain design endeavors, due to the potential identification of misleading reactions [10]. Nicotinamide adenine dinucleotide (NAD(H)) and nicotinamide adenine dinucleotide phosphate (NADP(H)), are the most wildly used cofactors in cell metabolism, being extensively examined for chemical processing applications [11]. These structurally similar molecules act as functional group transfer agents, being therefore consumed at the same rate of substrate consumption [12]. Moreover, the uncertainty of their usage in metabolic reactions has a major impact in metabolic engineering applications, affecting both predictions and strain design results [10]. When correctly characterized, enzyme modification by thorough structure redesign for cofactor specificity change can be undertaken, enabling the efficient processing of multiple desired biocatalytic transformations [13].

NAD(H) and NADP(H) are functionally equivalent cofactors used for storage and exchange of electrons in catalytic reactions. These cofactors are used by the majority of oxidoreductases, the largest class in Enzyme Commission. The only difference between these two molecules is a phosphate group in the adenine moiety, located on the opposite side from the chemically active nicotinamide moiety. Despite their apparent similarity, enzymes that use these cofactors tend to be specific for only one of them, enabling pathway regulation and chemical driving force maintenance by the cells, through heavy regulation of the levels of oxidized and reduced metabolic pools of NAD(P)(H) [14].

Strategies using sequence alignment and analysis of specific structural motifs such us Rossmann-folds, first identified by Rao and Rossmann in 1973 [15], have found limited success in identifying the responsible residues for NAD(H) / NADP(H) specificity based in these motifs, due to the variability of the residues present [16]. The Rossmann fold is a common structural motif found in many NAD(P)(H)-binding proteins, however, its short sequence and some inconsistencies found in the motif sequence question its reliability as a motif [17]. An approach using machine learning on protein sequences was able to correctly identify Rossmann-folds in sequence motifs, but still lacked accuracy in differentiating NAD(H) and NADP(H) specificity [18]. Also, a study using a dataset of NAD(P)(H)-binding enzymes has found that in nearly three-quarters of the analyzed dataset were represented at least twelve distinct locally conserved structural motifs for binding NAD(P)(H), having the remaining dataset no distinguishable motifs [16].

Carugo, et al. [19], in 1997, concluded that the only conserved structural feature in NADP(H) complexes is an Arginine residue present near the adenine moiety, interacting with the phosphomonoester through hydrogen bonds. For NAD(H) it was determined that the only clue for specificity identification was the presence of an Aspartate (or Glutamate) residue able to chelate the diol group of the ribose near the adenine. However, often these structural features are not present in structures bound to NAD(H) or NADP(H), hindering their correct characterization, and besides, structural information on the subject enzyme is required, which regularly is not the case.

Despite the efforts in identifying the cofactor specificity, very few studies go beyond pinpointing the specific residues Arginine and Aspartate and only for the phosphate moiety area. The best example is a study performed using ketol-acid reductoisomerases that showed that the presence of acidic residues at conserved phosphate binding positions are potential candidates of enzymes preferring NAD(H) [20]. Another problem is the fact that most studies are based on data often composed by small datasets or specific enzyme sub-classes, which can bias the results due to their sequence similarity, and are regularly characterized using visual interpretation or selection of positive cofactor change mutations [20]–[23]. This hurdle might also be explained by the fact that both NAD(H) and NADP(H) specific enzymes tend to share analogous sequence and structural motifs, due to their molecular similarity. To this day, metabolic engineering problems requiring cofactor specificity change, heavily rely on laborious and high cost approaches such us *in vivo* random mutagenesis and activity essays, which are not accurate nor efficient, wasting precious resources on multiple and undirected gene mutations that often do not produce the desired result. Previous studies have compiled extensive lists of such approaches [13], [24], [25].

Given this information, it becomes clear that the development of a transversal method for the accurate prediction of NAD(P)(H) cofactor specificity in uncharacterized enzymes is missing in this field. As the number of available protein structures increases in the Protein Data Bank (PDB) [26] (the October 24th, 2017, release has 134656 available structures, with 42572 being directly linked to Uniprot, being approximately 3000 bound to NAD(P)(H)), new approaches on this subject can be developed in order to enhance the accuracy of cofactor prediction while unveiling the responsible molecular determinants for cofactor specificity.

One of the most efficient approaches for performing such task is through the use of Machine learning, as these methods apply algorithms for mining large amounts of data in order to extract useful and unknown correlations, creating models able to predict and relate molecular descriptors to biological attributes [27].

Multiple successful approaches using machine learning and structural information have been implemented in the field of pharmacology, namely in pharmacokinetics and drug-discovery [28]–[32], suggesting that these methodologies may be adequate for cofactor specificity prediction.

In the present study the molecular determinants for NAD(P)(H) cofactor specificity were unveiled, using enzyme structural information. A comprehensive dataset of structures from enzymes using NAD(P)(H) as cofactors was build and processed using machine learning algorithms. The ensuing results were further analyzed to identify the responsible molecular factors for cofactor specificity.

These findings where successfully applied to enzymes not structurally characterized, using comparative modelling and a protocol was developed to automatically predict cofactor specificity. A webserver was also developed to allow a fast and easy-to-use access to the automatic prediction of NAD(P)(H) cofactor specificity of functionally uncharacterized enzyme sequences. Submissions can be performed for one or more aminoacid sequences in the FASTA format at http://services.itqb.unl.pt/cofactor-prediction/.

3.2 Methods

3.2.1 Structure analysis and CNRPM generation

The tool for generating the cofactor neighbor residue profile matrix (CNRPM) for each NAD(P)(H) bound enzyme structure was built using the python programming language. Each structure is automatically handled and the distances between each cofactor atom and the residue neighborhood are retrieved using the PDB module of the Biopython package [33].

Interactions between cofactor atom and neighbor residue are assembled in a matrix and outputted, in order to be processed by the machine learning algorithm.

3.2.2 CNRPM dataset extraction

All structures bound to one of the following ligand IDs: NAD/NAI/NAP/NDP, representatives of NAD$^+$/NADH/NADP$^+$/NADPH respectively, were sought after in the PDB and automatically retrieved and analyzed using the PDB module of the Biopython package. Entries whose enzyme or cofactor structure were incomplete or disrupted were discarded. In order to overcome the problem of overfitting/biasing the study with structure duplicates or point mutations of the same enzymes with different entry codes, a redundancy threshold was set and applied to the sequences coding the retrieved enzyme structures. The selected threshold was set to 95% similarity in 90% of the sequence length, allowing the removal of duplicates and point mutations of the same enzymes.

3.2.3 Machine learning

Machine learning was used for solving the classification problem in the form of supervised learning. Support vector machine, the selected method, was applied using the scikit-learn library for python

[34]. LIBSVM was the employed library and the radial basis function (RBF) was the chosen kernel function.

The developed CNRPM dataset was used as a training set and handled with the NumPy library for python [35]. Model performance was evaluated by measurements including accuracy, precision, Matthew's correlation coefficient (MCC) and area under curve of the receiver operating characteristics (AUC ROC). Accuracy refers to the closeness of a measured value to a standard or known value, precision refers to the closeness of two or more measurements to each other. MCC measures the prediction quality, taking into account over- and under- predictions and giving a complementary measure of the prediction performance[36]. MCC of 1 means a perfect prediction, and 0 denotes a completely random prediction. The receiver operating characteristic (ROC) curve [37], plots true positive rate on the y-axis against the false positive rate on the x-axis. The normalized area under curve of the receiver operating characteristics (AUC ROC) states a perfect prediction if the AUC value is 1, and a random guess if the value is 0.5.

3.2.4 Comparative modeling for structure analysis

Homology models were created using Modeller [38] and the modeller package for python, where sequence similarity search for template selection was performed using the Smith-Waterman local alignment [39], [40] in a local database composed by structures from PDB bound to one of the cofactors NAD(P)(H). Structural similarity evidencing a suitable template was assume when two proteins share a sequence identity above 25%

The structure of the Cofactor was correctly allocated in the modelled structures by allowing Modeller to transfer these molecules from the template to the modelled structure.

3.2.5 NiCofactor tool construction

The created tool for allowing high throughput NAD(P)(H) cofactor specificity prediction was built using the python programming language. For each sequence in the FASTA format used as input, the tool initiates an individual project. The tools for generating CNRPMs and performing machine learning were also integrated in NiCofactor. Results are outputted by attributing to each analyzed sequence a cofactor prediction and subsequent prediction score.

3.2.6 NiCofactor result validation dataset

Curated information on cofactor, cofactor specificity, EC number, organism, sequence, literature and source information on enzymes using NAD(P)(H) were retrieved automatically from brenda-enzymes using SOAPpy, a tool for building SOAP clients and servers, implemented in python [41].

3.3 Results and discussion

In the present chapter we performed a comprehensive study, using big data on protein structural information and machine learning, in order to unveil the molecular determinants of cofactor specificity in enzymes using NAD(P)(H) as cofactors. Previous attempts pinpointed a sequence motifs and some prevalent residues near the 2'-phosphate, either by sequence analysis or successful mutations for cofactor specificity change, using random mutagenesis [13], [15], [19]–[25]. However, to this day, and our knowledge, there is still to be made a transversal structural study on the interactions between cofactor binding pocket residues and cofactor atoms, using a large dataset of enzyme structures. Such task required large amounts of data to be retrieved and analyzed, being such analysis unfeasible without the use of machine learning algorithms that can find seemingly undetectable patterns in big data for further use in the accurate prediction of cofactor specificity for less characterized enzymes.

3.3.1 Cofactor neighbor residue profile matrix (CNRPM) development

Characterizing structural information can be a challenging task due to the overwhelming amount of information associated with the structure of a protein. Our main focus was to retrieve all possible interactions between each cofactor atom and the nearest residues in the binding pocket. With that in mind we developed a tool that, given a characterized structure bound with NAD(P)(H) (in the PDB format), automatically returns a matrix of interactions between each cofactor atom and the surrounding amino-acid residues, at a distance of 6 Å. By ignoring the atoms related to the phosphate in the adenosine moiety of NADP(H), we were able to create similar cofactor neighbor residue profile matrices (CNRPM) for both NAD(H) and NADP(H) cofactors, which is crucial to a well performing machine learning method. Figure 3.1 depicts the cofactor neighbor residue profile matrix building process.

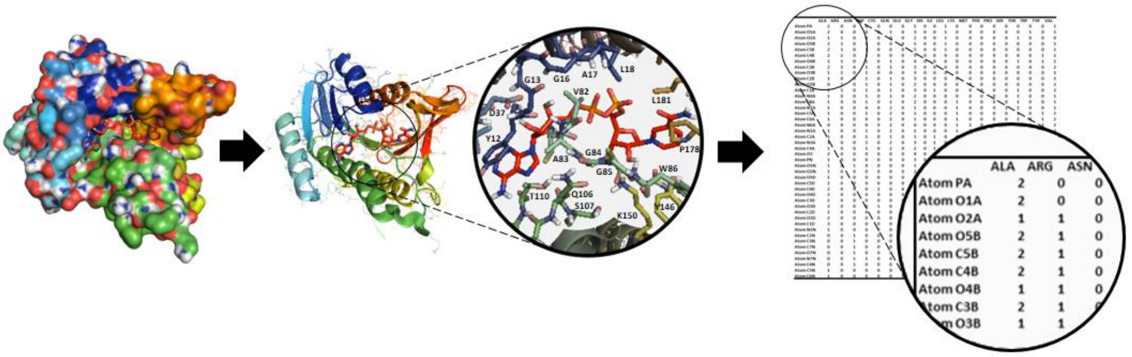


Figure 3.1 - Cofactor neighbor residue profile matrix generation. Starting with the whole structure, the tool pinpoints the location of the cofactor as a reference and proceeds to register the position of its atoms. For each cofactor atom, the tool investigates the surrounding residues and catalogues those within 6 Å. The complete process is performed automatically. The end result encompasses a matrix of interactions between each cofactor atom and neighboring residues. The presented enzyme structure depicts a Dihydropteridine Reductase bound to NAD$^+$ from Rat liver, with the EC 1.5.1.34. PDB id: 1DIR.

In these CNRPM, where each line refers to a cofactor atom (44 atoms) and each column refers to one of the twenty natural amino-acids, each value refers to the number of residues found. If, within the surroundings of an atom, a specific residue is not present, the value of that interaction is set to 0 (zero) in the matrix. The final product of the developed tool, the CNRPM, is a 20x44 matrix encompassing 880 interaction values.

3.3.1.1 Building a comprehensive and representative CNRPM dataset

With the intent of applying the developed tool in the construction of an accurate and representative dataset of CNRPMs, for unveiling the molecular determinants of cofactor specificity, a database of enzyme structures bound to NAD(P)(H) was assembled. To do so, we retrieved (in January 13th 2016) all enzyme structures bound to one of the cofactors NAD(P)(H) from the PDB and analyzed them. The total amount of structures collected was 2742, from which 148 were discarded due to incompleteness. With the removal of protein sequence redundancy, the final dataset encompassed 921 structures, being 491 structures bound to NAD(H) and 430 to NADP(H). Once the database was assembled and validated, the developed tool was applied to all structures and a CNRPM was retrieved for every entry.

3.3.2 CNRPM dataset analysis and processing using Machine learning

Having built a large representative dataset of 921 CNRPMs, we used support vector machine (SVM) algorithms to attribute cofactor preference based on the CNRPMs, while evaluating the performance of the method. The SVM training algorithm works by building a model, with categorized training examples, such as the CNRPMs (which are categorized as belonging to NAD(H) or NADP(H)), and representing them as points in a high-dimensional hyperplane, separated by category and divided by a clear gap between them. This allows the algorithm to assign a category to uncategorized new examples, based on the side of the hyperplane they fall. Performance is assessed by measuring how fine the division of categories is achieved [42].

By applying this algorithm to our CNRPMs dataset as a training set, an SVM model was created whose evaluation and performance parameters can be found in table 3.1. The created model achieved an accuracy of 96.2%, being able to correctly classify 886 CNRPMs as corresponding to NAD(H) or NADP(H) cofactors, with a precision of 96.03% and a Matthews correlation coefficient (MCC) of 0.92. The computed area under the receiver operating characteristic curve (AUC ROC) coefficient is 0.96. The confusion matrix displayed in table 3.1 evidences the high sensitivity and specificity of the model, with similar misclassification values in both NAD(H) and NADP(H) CNRPM.

Table 3.1 - Evaluation and performance parameters of the created SVM model. Accuracy, precision, MCC (Mathews correlation coefficient) and AUC ROC (area under the receiver operating characteristic curve) values (top) display the overall performance of the mode, indicating a well performing model. The Confusion matrix (bottom) evaluates sensitivity and specificity of the model.

	Accuracy	**Precision**	**MCC**	**AUC ROC**
SVM model	96.20%	96.03%	0.92	0.96

Real cofactor			
NAD(H)	**NADP(H)**		
474	17	**NAD(H)**	**Predicted cofactor**
18	412	**NADP(H)**	

These results put in evidence that the type and number of residues present in the cofactor binding site have a crucial role in the specification of cofactor preference in the enzyme. Such results also demonstrate the possibility to predict/indicate cofactor preference in an enzyme by analyzing its cofactor neighbor residue profile using these methodologies.

3.3.2.1 SVM feature weights extraction and interpretation

The SVM model training works by attributing weights to features in the dataset (in this case a feature is a cofactor atom-residue interaction), allowing the correct separation of the instances in the hyperplane. Such separation is what enables the algorithm to classify a CNRPM as originated from an enzyme bound to NAD(H) or NADP(H). The extraction and interpretation of such metrics are of great importance in the identification of the crucial interactions between residue and cofactor atoms, and should allow us to exactly pinpoint the set of relations in the CNRPM responsible for providing the cofactor preference to an enzyme. The extracted data, composed by 880 features and their respective weight in the SVM model, are presented in table A1 of the appendix. The highest extracted weight values correspond to 0.44991 for NADP(H) and 0.23609 for NAD(H). Despite the large amount of features, feature weight values from both NADP(H) and NAD(H) decrease rapidly from the heaviest values, leveling out in lighter features. This indicates that, despite the contribution of all features to the classification of the CNRPMs, some relations have a more significant role in classifying cofactor preference than others.

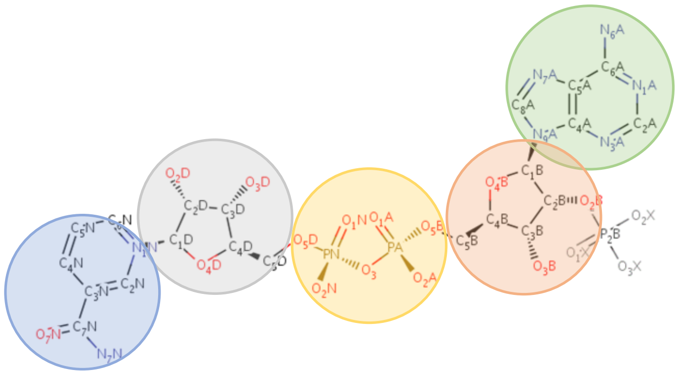
When analyzing the results, it was possible to observe that atoms from all parts of the cofactor structure contribute to specificity, despite the only difference between both cofactors being the presence of a phosphate molecule in the ribose from the adenine moiety. In fact, the fifteen heaviest features for both cofactors encompass atoms from adenine, ribose from adenosine, phosphates, ribose from nicotinamide ribose and nicotinamide. The contributions of Aspartate and Arginine for NAD(H) and NADP(H) specificity, respectively, and already reported in several publications, are also observed in the retrieved data. The presence of these residues near the adenosine atoms represent some of the heaviest features. However, claims that attribute vital importance of these residues in cofactor specificity are probably simplistic, resulting from studies using small datasets or the manual (case-by-case) observation of data. Table 3.2 displays the fifty heaviest features for each cofactor along with the respective weight. A color scheme helps locating the area of the cofactor correspondent to each atom.

Table 3.2 - SVM model feature weight distribution for NAD(H) (left) and NADP(H) (right). Feature weight is distributed in a decreasing order, starting from the heaviest. Columns depict the type of atom, residue (AA) and feature weight. Feature weights are divided into two sub columns for each cofactor. Colored cell in the atom columns represent the different areas composing the cofactor structure.

NAD(H)						NADP(H)					
Atom	AA	Weight	Atom	AA	Weight	Atom	AA	Weight	Atom	AA	Weight
C8A	ASP	0.236	C4B	ARG	0.144	O2B	ARG	0.450	C4N	LEU	0.145
O4B	SER	0.214	N3A	ASP	0.143	O2B	SER	0.265	C4A	ASN	0.137
C4B	ASP	0.212	O2B	LEU	0.142	O2B	LYS	0.262	N9A	TYR	0.137
C5B	ASP	0.212	O3B	GLU	0.141	C2B	ARG	0.229	O2D	ALA	0.135
O5B	ASP	0.202	N9A	LEU	0.139	O3	GLY	0.223	O2A	ALA	0.134

C5A	LYS	0.199	C1B	ALA	0.137	O3B	ALA	0.218	O5B	SER	0.133
O2D	VAL	0.193	C4D	TYR	0.135	C4D	ASN	0.217	N1A	PRO	0.132
C2B	GLU	0.179	O1N	ARG	0.134	O1A	ALA	0.186	N7N	ASN	0.129
N6A	ALA	0.179	C3N	GLU	0.128	O2B	GLY	0.180	O1A	LYS	0.127
C5B	TRP	0.177	O5D	ALA	0.128	N1A	SER	0.168	N7N	ALA	0.127
O1N	LEU	0.176	N7N	ILE	0.126	C5B	LEU	0.164	O3B	LYS	0.124
C2B	GLY	0.174	N9A	GLU	0.125	C1B	ARG	0.164	C5D	ALA	0.123
N7A	PHE	0.165	O1A	ARG	0.125	O3D	TYR	0.161	C2B	SER	0.123
O2N	LEU	0.162	O3B	ILE	0.125	O7N	CYS	0.159	C4N	GLY	0.119
N7N	ASP	0.157	C8A	MET	0.122	PN	GLU	0.159	C8A	TYR	0.117
O2N	PHE	0.157	O5D	THR	0.122	C2N	ILE	0.158	O3	ASN	0.116
O3B	PHE	0.155	O1N	ASN	0.121	C5A	TYR	0.157	O4D	THR	0.115
C5B	ALA	0.153	O7N	ARG	0.120	O2N	ASP	0.153	C4A	ALA	0.114
O3B	GLY	0.152	N3A	LYS	0.120	C3N	LYS	0.150	C4B	LEU	0.114
O2B	GLU	0.152	N1A	PHE	0.120	O5D	ILE	0.149	N1N	ASP	0.114
O1N	TYR	0.152	N9A	ILE	0.120	C4D	THR	0.148	C5B	CYS	0.113
N3A	TYR	0.151	C3B	ILE	0.118	N7A	TYR	0.148	O4B	GLY	0.112
N9A	ASP	0.150	C1B	SER	0.118	C2B	LYS	0.147	C2B	THR	0.112
O5D	GLN	0.149	C5A	PRO	0.117	O1A	SER	0.146	O7N	HIS	0.111
C2B	PRO	0.148	C3D	HIS	0.117	C6A	VAL	0.145	C8A	SER	0.110

🟩	Adenine
🟧	Ribose (adenine)
🟨	Phosphates
⬜	Ribose (nicotinamide)
🟦	Nicotinamide



In Table 3.2, not only cofactor atoms from the entire cofactor structure are present, but also a large majority of the 20 natural amino acids residues are present in features from both cofactors. In the case of NAD(H), besides Aspartate, also Glutamate, Alanine, Leucine, Phenylalanine, Arginine and Isoleucine residues are frequently present in the displayed features, dispersed in interactions with atoms from the entire NAD(H) structure, being Cysteine the only amino acid residue not present in the first fifty features. In the case of NADP(H), again the most important interactions occur in atoms belonging to the adenosine moiety, with Arginine residues near the atom O2B being the heaviest feature, possibly due to the presence of the phosphate connected to that atom in NADP(H). Serine, Lysine, Glycine, Alanine, Asparagine and Tyrosine residues are the most frequent aminoacid residues present in the first features, being absent from this group Tryptophan, Phenylalanine, Glutamine and Methionine residues.

3.3.3 NiCofactor cofactor specificity prediction tool development

With the important intent of developing a robust and high throughput method for NAD(P)(H) cofactor specificity prediction for enzymes whose structure is yet to be characterized, we decided to use comparative modelling methods within our pipeline. These methods will not only allow processing newly sequenced enzymes or organisms, but also cope with the large existing gap between available sequences in Uniprot (93 million) and structures in PDB (almost 135 thousand, with only 42572 being directly linked to Uniprot as of October, 2017). To do so, we developed a method that implements functions for comparative modelling of protein structures using Modeller [38], a software that performs modeling by satisfaction of spatial restrains, through sequence alignment of the target sequence and known related structure templates. Through the integration of the developed methods with the resulting SVM model, we created a tool that automatically performs cofactor preference prediction. With only the input of an amino acid sequence, a machine learning analysis of the modeled structural environment around the cofactor is performed. Figure 3.2 represents the pipeline developed within the built framework that enables the prediction of cofactor specificity. The developed framework is implemented in a computational tool and project submissions are enabled via the freely available online webserver http://services.itqb.unl.pt/cofactor-prediction/.

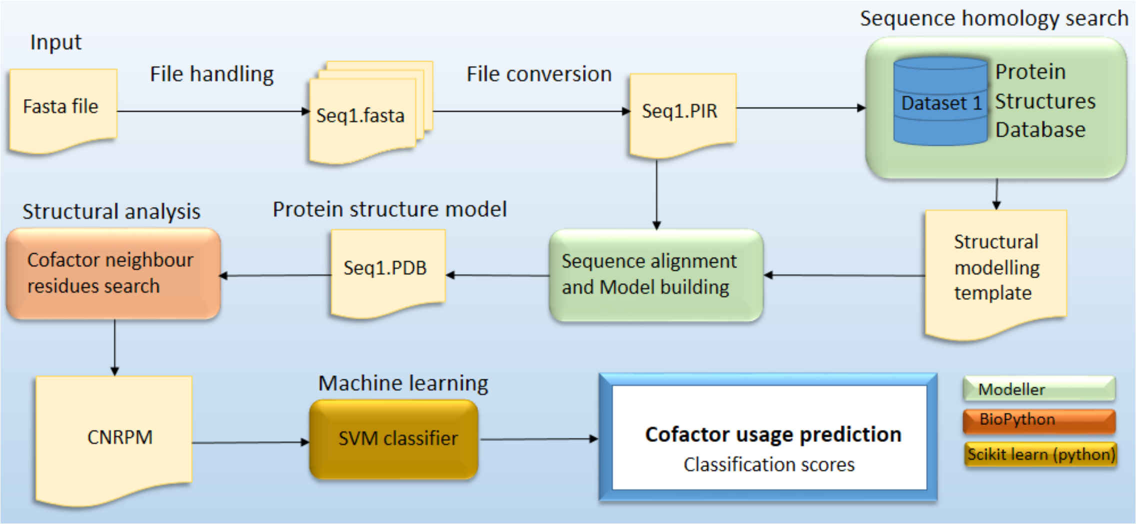


Figure 3.2 - Developed framework pipeline. The displayed planes depict the sequence of events necessary for cofactor prediction. Starting from the left top, a Fasta file composed by the target aminoacid sequences is inputted in the system where they are structurally modelled, analyzed and classified.

3.3.3.1 Validation of NiCofactor cofactor specificity prediction tool using curated information

After successfully developing an accurate SVM model using structural information as a training set and integrating it in a cofactor prediction tool, model and tool validation is still a requirement. In order to validate the developed method and the machine learning model, a dataset with curated information on enzyme specificity was constructed. For that, Brenda-enzymes [9] was used and curated information on cofactor, cofactor specificity, EC number, organism, sequence, literature and source were retrieved. Firstly, the database was filtered for enzyme entries with NAD$^+$, NADH, NADP$^+$ or NADPH as cofactor, subsequently the cofactor commentary field was filtered for expressions indicating high cofactor specificity, such us, "absolute specificity", "specific", "totally specific", "dependent on", "strict", "no activity" or "required". This step enabled us to create a dataset of 404 distinct aminoacid residue sequences of different enzymes with high cofactor specificity experimentally determined, originated from a combination of 198 EC numbers and 180 organisms. From the total amount of enzyme aminoacid sequences, 189 are specific for NAD(H) and 215 for NADP(H). With the retrieved information present on the dataset, the aminoacid sequences encompassed in the dataset were arranged and displayed in Fasta format, in a single Fasta file. This file was then uploaded and processed in the developed tool NiCofactor and predictions of NAD(P)(H) cofactor specificity were performed.

When analyzing the results obtained from the developed tool it was possible to observe that from the total 404 sequences analyzed, composing the dataset of curated information, the developed tool performed a cofactor prediction for 327 (81%) as around 11% (45) of the enzymes analyzed had their structure characterized and approximately 70% (282), despite not having their structure characterized, were found to have a suitable structural template, enabling structure inference by homology modeling. For 19% (77) of the enzymes analyzed, no structural template was found, impairing the possibility of a cofactor prediction being performed by the tool. The overall accuracy of the tool predictions was 83.5%, which is by itself an extremely exciting result when taking into account that only an aminoacid residue sequence is given as input, with the tool automatically performing the structural analysis. Nonetheless, by further analyzing the machine learning model prediction output, it is possible to retrieve the predictions probability, which is an estimate from the model on how probable the prediction is correct. When plotting the prediction probability results we observed that the accuracy of the model tends to increase with the prediction probability, which opens the possibility of establishing a probability threshold that should improve the framework predictive capabilities. Figure 3.3 plots the values of prediction probability with the corresponding sequence distribution and prediction accuracy of the performed predictions.

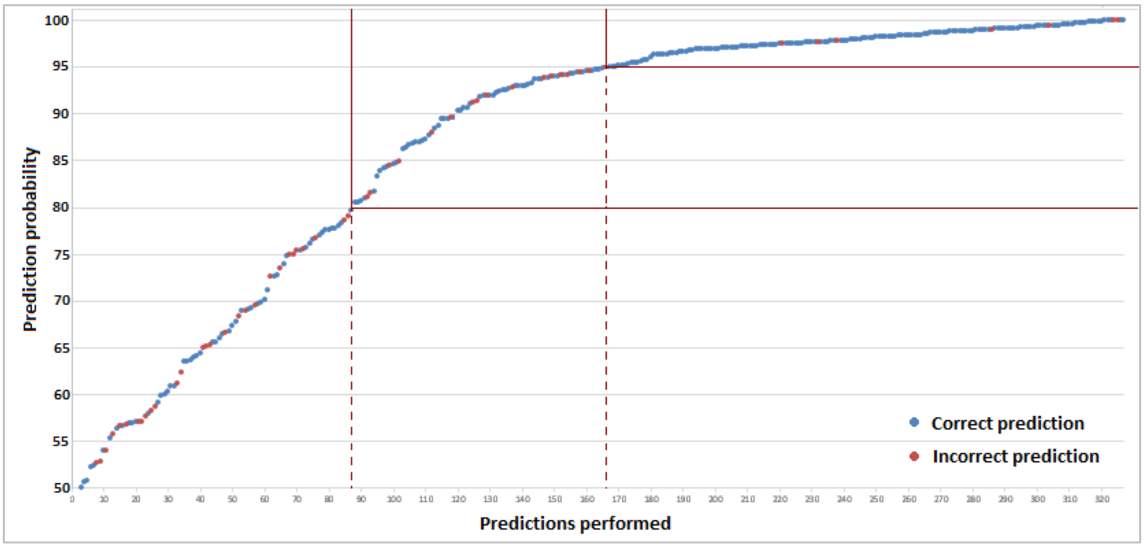
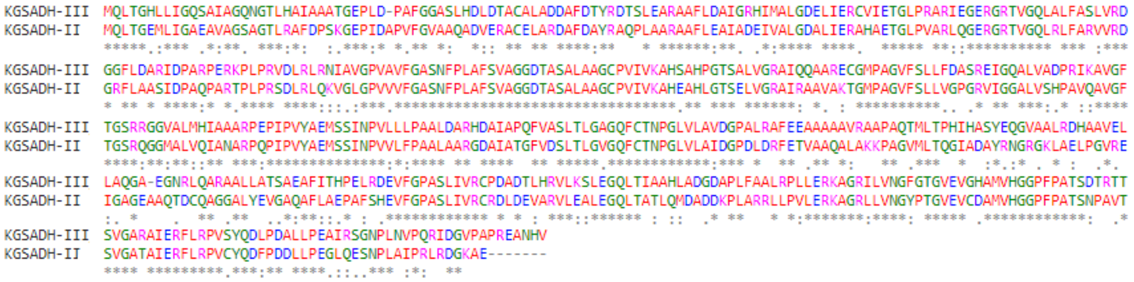


Figure 3.3 - SVM model validation graph. Plotted are all predictions performed (x axis) and the correspondent SVM model outputted prediction probability (y axis). Colors of the points refer correct (blue) and incorrect (red) predictions performed by the tool. Red lines delineate prediction probability of 80 and 95 percent, and evidence the amount of predictions encompassed in those probabilities.

From figure 3.3 we can notice that the vast majority of predictions made by the SVM model, within the developed tool, have a very high probability score, with nearly 50% of the analyzed sequences having a cofactor prediction probability of at least 95%, according to the model. In fact, 73.4% (240) of the outputted predictions have a prediction probability above 80%. This results indicate that most of the prediction performed by the developed SVM model have a high probability of being correct. It is also possible to observe that the accuracy of the predictions made increases with the prediction probability score outputted by the model, which validates the prediction probability score. When the outputted prediction has a probability of at least 80%, the accuracy of the predictions increases to from 83.5% to 90%, and when the prediction probability surpasses 95%, model accuracy is 96%. The presented results validate the developed tool for performing NAD(P)(H) cofactor preference predictions on enzymes, using only the enzyme's aminoacid residue sequence as input, which enables the prediction of cofactor preference in newly sequenced enzymes, or enzymes whose structures are yet to be characterized. To improve the prediction accuracy of the developed tool, a prediction probability threshold of 80% was set, which means that if the prediction probability of an analyzed sequence is, at least, 80%, the prediction is accepted as correct.

3.3.3.2 NiCofactor sensitivity analysis with case studies

In order to demonstrate the performance sensitivity of the developed tool, two experimentally characterized case studies encompassing homologue enzymes using distinct cofactors, were further analyzed. First, the case of *Azospirillum brasiliense*'s α-Ketoglutaric Semialdehyde Dehydrogenase Isozymes (KGSADH) is presented. According to Watanabe *et al.* [43] in *A. brasiliense*, KGSADH is involved in the conversion of α-ketoglutaric Semialdehyde to α-Ketoglutarate in an alternative pathway of L-arabinose metabolism. In his study it is described that this bacterium encodes for two different KGSADH isozymes, D-glucarate/D-galactarate-inducible KGSADH-II and hydroxy-L-proline-inducible KGSADH-III with significantly similar sequences. After physiological characterization, they revealed that KGSADH-II and KGSADH-III showed similar high substrate specificity for α-ketoglutaric semialdehyde and different cofactor specificity, being KGSADH-II, NAD^+ dependent and KGSADH-III, $NADP^+$ dependent. Figure 3.4 illustrates the sequence alignment between KGSADH-II and KGSADH-III, where it is possible to observe the sequence similarity between these enzymes. KGSADH-II and KGSADH-III have a sequence identity of 62.41%, with 332 identical residues in an alignment length of 532.



```
KGSADH-III  MQLTGHLLIGQSAIAGQNGTLHAIAAATGEPLD-PAFGGASLHDLDTACALADDAFDTYRDTSLEARAAFLDAIGRHIMALGDELIERCVIETGLPRARIEGERGRTVGQLALFASLVRD
KGSADH-II   MQLTGEMLIGAEAVAGSAGTLRAFDPSKGEPIDAPVFGVAAQADVERACELARDAFDAYRAQPLAARAAFLEAIADEIVALGDALIERAHAETGLPVARLQGERGRTVGQLRLFARVVRD
            *****.:*** .*:**. ***:*:  :.***:* *.** *:  *:: ** ** ****:**   * ******:**. .*:**** ****.  ***** **::********** *** :***

KGSADH-III  GGFLDARIDPARPERKPLPRVDLRLRNIAVGPVAVFGASNFPLAFSVAGGDTASALAAGCPVIVKAHSAHPGTSALVGRAIQQAARECGMPAGVFSLLFDASREIGQALVADPRIKAVGF
KGSADH-II   GRFLAASIDPAQPARTPLPRSDLRLQKVGLGPVVVFGASNFPLAFSVAGGDTASALAAGCPVIVKAHEAHLGTSELVGRAIRAAVAKTGMPAGVFSLLVGPGRVIGGALVSHPAVQAVGF
            * ** * ****:* *.**** ****:::.:***.*********************************.** *** ******: *. : **********.. .* ** ***:.* ::****

KGSADH-III  TGSRRGGVALMHIAAARPEPIPVYAEMSSINPVLLLPAALDARHDAIAPQFVASLTLGAGQFCTNPGLVLAVDGPALRAFEEAAAAAVRAAPAQTMLTPHIHASYEQGVAALRDHAAVEL
KGSADH-II   TGSRQGGMALVQIANARPQPIPVYAEMSSINPVVLFPAALAARGDAIATGFVDSLTLGVGQFCTNPGLVLAIDGPDLDRFETVAAQALAKKPAGVMLTQGIADAYRNGRGKLAELPGVRE
            ****:**:**::** ***:**************:*:**** ** ****  ** *****.************:*** *  ** .** *:   ** .***  *  :*.:* . * :  .*.

KGSADH-III  LAQGA-EGNRLQARAALLATSAEAFITHPELRDEVFGPASLIVRCPDADTLHRVLKSLEGQLTIAAHLADGDAPLFAALRPLLERKAGRILVNGFGTGVEVGHAMVHGGPFPATSDTRTT
KGSADH-II   IGAGEAAQTDCQAGGALYEVGAQAFLAEPAFSHEVFGPASLIVRCRDLDEVARVLEALEGQLTATLQMDADDKPLARRLLPVLERKAGRLLVNGYPTGVEVCDAMVHGGPFPATSNPAVT
            :. *    .  ** .**  ..*:**::.* : .************ * * : ***::****** : ::  .* **   * *:*******:****: ***** .************:  .*

KGSADH-III  SVGARAIERFLRPVSYQDLPDALLPEAIRSGNPLNVPQRIDGVPAPREANHV
KGSADH-II   SVGATAIERFLRPVCYQDFPDDLLPEGLQESNPLAIPRLRDGKAE-------
            **** *********.***:** ****.::..*** :*:  **
```

Figure 3.4 - KGSADH II and KGSADH III aminoacid sequence alignment. KGSADH III (top) and KGSADH II (bottom) aminoacid sequence alignment.

The second case presented regards to two alkyl alcohol dehydrogenase (ADH) genes from the long-chain alkane-degrading strain *Geobacillus thermodenitrificans* NG80-2 characterized by Liu, et al. [44]. Both ADH1 and ADH2 are able to oxidize a broad range of alkyl alcohols up to at least C_{30}, as well as 1,3-propanediol and acetaldehyde, and share a sequence identity of 26%. For either enzyme, both NAD^+ and $NADP^+$ can be used as electron acceptor. However, NAD^+ is the preferred cofactor for ADH1, while $NADP^+$ is the preferred cofactor for ADH2.

With the presented information we went on to perform cofactor prediction and assess the capability of the developed tool to predict the cofactor preference of such similar enzymes. With none of the structures of the analyzed enzymes characterized, the framework applied structure modelling in order to perform a prediction. After processing all sequences using NiCofactor for performing cofactor predictions, the resulting output was analyzed. In it, we were able to verify that the tool correctly classified the analyzed enzymes, being KGSADH-II classified as NAD(H) binding with a prediction probability of 99.2% and KGSADH-III as NADP(H) with 64.7% probability, whereas for the ADH genes, ADH1 was classified as NAD(H) specific with a prediction probability of 80.6% and ADH2 predicted as NADP(H) specific with 95.7% of probability.

These results demonstrate the robustness of the developed tool in correctly attributing NADP(P)(H) cofactor preference to enzymes, using the aminoacid sequence as input. The results from the performed predictions are displayed in table 3.3. In the case of KGSADH, possibly due to their similarity, the selected structure template for both enzymes was the same, 1EZ0.pdb, a NADP$^+$ dependent Aldehyde dehydrogenase from *Vibrio harveyi*, characterized by Ahvazi *et al*. [45]. This enzyme has a sequence similarity of 48% with KGSADH-II and 47% with KGSADH-III, being its structure characterized with NADP$^+$ in the binding pocket. Regardless of the type of bound cofactor in the template enzyme structure, the developed method was still able to correctly classify cofactor preference in the subject enzymes, being the prediction with higher probability from the opposite cofactor. ADH1 structure was modelled using an alcohol dehydrogenase structure from *Thermotoga maritima* (PDB: 1O2D), with a sequence identity of 37%, while ADH2 model template was a butanol dehydrogenase also from *Thermotoga maritma* (PDB: 1VLJ), with 48% sequence identity.

Table 3.3 - Cofactor specificity prediction. KGSADH II and KGSADH III from *Azospirillum brasiliense* cofactor specificity prediction analysis show the predicted cofactor and associated probability. ADH1 and ADH2 from *Geobacillus thermodenitrificans* NG80-2 cofactor specificity prediction analysis show the predicted cofactor and associated probability. Template information is also displayed with PDB ID and crystalized cofactor, as well as subject and template aminoacid sequence alignment identity percentage.

Fasta_ID	Predicted cofactor	Probability	Selected template	Alignment identity %
KGSADH-II	NAD(H)	0.992	1EZ0.PDB (NAP)	48
KGSADH-III	NADP(H)	0.647	1EZ0.PDB (NAP)	47
ADH1	NAD(H)	0.806	1O2D.PDB (NAP)	37
ADH2	NADP(H)	0.997	1VLJ.PDB (NAP)	45

The fact that homologue enzymes are usually specific for only one of the cofactors impairs a deeper analysis of case studies with homologues that use different cofactors. Nonetheless, the studied cases still present a good indicator of the performance sensitivity achieved. These case studies also help demonstrating that enzymes within the same environment, and with very similar functions and sequences, do not necessarily use the same cofactors for catalysis, which is a common assumption when annotating enzymes using sequence homology information.

3.3.3.3 NiCofactor tool usage for assessing cofactor engineering mutations impact in specificity

Another important application of the developed tool might reside in its ability of assessing the impact of point and combined mutations in cofactor specificity change strategies. We strongly believe that predicting the outcome of such strategies hugely improves the overall process efficiency and leverages the whole metabolic engineering field. In order to prove such capability a group of cofactor engineering studies published by Khoury and coworkers [13] and extending the work from Marohnic, *et al.* [25] was thoroughly analyzed. To perform such analysis, each enzyme was sought after in Uniprot and the corresponding aminoacid sequence retrieved, when available. The wild-type and mutant sequences were reproduced *in silico* and, subsequently, analyzed in the developed tool. Not all experiments were able to be reproduced due to the lack of sequence or mismatches between the retrieved sequence and mutations reported in the analyzed experiment. From a total of 27 cofactor engineering studies compiled, only for 18 was it possible to retrieve the aminoacid residue sequences, with a combined number of 35 mutations experimentally characterized. The subject studies and correspondent cofactor specificity analysis are displayed in table 3.4.

Regarding the 18 wild-type sequences analyzed, it was possible to retrieve structural information for all, either due to it being already characterized or through homology modelling. As to cofactor predictions, NiCofactor performed a prediction above the probability threshold for 14 (~78%) with 100% accuracy, being the cofactor specificity of all predictions correctly attributed.

Table 3.4 - Wild-type enzymes with cofactor predictions above the probability threshold. Displayed are the organism, enzyme name, number of mutations implemented in the analyzed studies, wild-type cofactor specificity, predicted cofactor specificity and probability score.

Organism	Enzyme	Mutants constructed	Wild-type specificity	Predicted cofactor	Cofactor probability	Ref
Candida tenuis	Xylose reductase	7	NADPH	NADP(H)	88.41	[46], [47]
Corynebacterium	2,5-diketo-D-gluconic acid	4	NADPH	NADP(H)	98.67	[48], [49]
Escherichia coli	Ketol acid reductoisomerase	4	NADPH	NADP(H)	99.05	[50]
Neurospora crassa	Nitrate reductase	1	NADPH	NADP(H)	95.76	[51]
Pichia stipitis	Xylose reductase	2	NADPH	NADP(H)	87.18	[52], [53]
Pseudomonas fluorescens	p-hydroxybenzoate hydroxylase	1	NADPH	NADP(H)	94.87	[54]
rattus norvegicus	Cytochrome p450 reductase	1	NADPH	NADP(H)	97.39	[55]
sinorhizobium morelense	1,5-anhydro-d-fructose	1	NADPH	NADP(H)	97.83	[56]
Escherichia coli	Isocitrate dehydrogenase	1	NADP$^+$	NADP(H)	95.01	[57]
Spinacia oleracea	Nitrate reductase	1	NADH	NAD(H)	84.69	[58]
Gluconobacter oxydans	Xylitol dehydrogenase	1	NAD$^+$	NAD(H)	96.55	[59]
Pichia stipitis	Xylitol dehydrogenase	1	NAD$^+$	NAD(H)	97.95	[60]
Thermus thermophilus	Isopropylmalate dehydrogenase	1	NAD$^+$	NAD(H)	98.26	[61]
Tramitichromis intermedius	Leucine dehydrogenase	1	NAD$^+$	NAD(H)	97.74	[62]

.

Concerning the mutations performed in the cofactor reversal engineering studies for the 14 above displayed enzymes, a total of 27 mutants were constructed, with 3 enzyme from *Candida tenuis*, *Corynebacterium* and *Escherichia coli* being responsible for 15 of the mutants produced.

After being analyzed with NiCofactor, a prediction above the probability threshold was made for 20 out of the total 27 (74.1%) mutation experiments analyzed, corresponding to 10 different enzymes. The performed mutations and respective predictions are displayed in table 3.5, along with the organisms and enzyme identification.

Table 3.5 - Performed mutations and respective cofactor predictions. Displayed are also the organism and enzyme, along with mutant cofactor usage and the specification of the mutation performed. Mutations are represented by the original aminoacid, in single letter code, followed by the aminoacid position in the sequence and the mutant aminoacid, also in a single letter coding format. Multiple mutations occurring in a single mutant are separated by a slash ('/').

Organism	Enzyme	Mutant specificity	Mutations	Predicted cofactor	Ref
Candida tenuis	Xylose reductase	NADH	K274R	NADP(H)	[46]
Candida tenuis	Xylose reductase	NADH	K274G	NADP(H)	[47]
Candida tenuis	Xylose reductase	NADH	N276D	NADP(H)	[47]
Candida tenuis	Xylose reductase	NADH	K274R/N276D	NADP(H)	[47]
Corynebacterium	2,5-diketo-D-gluconic acid	NADH	K232G	NADP(H)	[48]
Corynebacterium	2,5-diketo-D-gluconic acid	NADH	R235G	NADP(H)	[48]
Corynebacterium	2,5-diketo-D-gluconic acid	NADH	R238H	NADP(H)	[48]
Corynebacterium	2,5-diketo-D-gluconic acid	NADH	F22Y/RS233T/R235E/A272G	NADP(H)	[49]
Escherichia coli	Ketol acid reductoisomerase	NADH	R68D	NADP(H)	[50]
Escherichia coli	Ketol acid reductoisomerase	NADH	K69L	NADP(H)	[50]
Escherichia coli	Ketol acid reductoisomerase	NADH	K75V	NADP(H)	[50]
Escherichia coli	Ketol acid reductoisomerase	NADH	R76D	NADP(H)	[50]
Pichia stipitis	Xylose reductase	NADH	K270M	NADP(H)	[52]
Pichia stipitis	Xylose reductase	NADH	K270S/S271G/N272P/R276F	NAD(H)	[53]
Pseudomonas fluorescens	p-hydroxybenzoate hydroxylase	NADH	R33S/Q34R/P36R/D37A/Y38E	NADP(H)	[54]
sinorhizobium morelense	1,5-anhydro-d-fructose	NADH	A13G/S33D	NADP(H)	[56]
Spinacia oleracea	Nitrate reductase	NADPH	E864S/F876R	NAD(H)	[58]
Gluconobacter oxydans	xylitol dehydrogenase	NADP$^+$	D38S/M39R	NADP(H)	[59]
Pichia stipitis	Xylitol dehydrogenase	NADP$^+$	D207A/I208R/F209S/N211R	NADP(H)	[60]
Tramitichromis intermedius	Leucine dehydrogenase	NADP$^+$	D203A/I204R/D210R	NADP(H)	[62]

In the displayed table we can observe that from the mutants obtained for the ten different enzymes analyzed, only for 4 enzymes was NiCofactor able to correctly predict cofactor specificity. However, when further exploring the results achieved in the analyzed studies, it became clear that in many experiments a complete conversion of cofactor specificity was not achieved. In fact, from the 16 mismatches between mutant cofactor specificity and the tool output predictions, only p-hydroxybenzoate hydroxylase from *Pseudomonas fluorescens* was confirmed in the literature has having completely reverted cofactor specificity, due to the mutation of 5 aminoacids in the cofactor binding spot [54]. For all other mismatch cases, literature results showed that successful mutation reports were based on marginal decreases of *Km* values or increase in *Kcat* values for the desired cofactor, despite the fact that the native cofactor preference remained, or was not measured. This analysis showed that the reported mutations were only able to marginally enhance the desired cofactor acceptance, not disrupting native cofactor specificity, promoting, at the best, cofactor promiscuity. A reason for such minor changes observed in mutant cofactor specificity might be explained by the fact that the vast majority of the constructed mutants have only one residue mutated, as it has been shown that multiple simultaneous mutations have often to be performed in order to effectively change cofactor specificity, hindering this challenging task [63].

Correctly classified predictions were found to have completely reverted specificity or largely decreased affinity for one of the cofactors, increasing the affinity of the other. These were the cases of xylose reductase from *Picchia stipitis* [53], xylitol dehydrogenase from *Gluconobacter oxydans* [59], xylitol dehydrogenase from *Pichia stipitis* [60] and leucine dehydrogenase from *Tramitichromis intermedius* [62], all with multiple mutations implemented in the wild-type enzyme sequence. These results help emphasizing the utility of the developed framework in correctly predicting a clear, and substantial, specificity transfer from one cofactor to the other.

3.4 Conclusions

Molecular characterization of NAD(P)(H) cofactor specificity is, to this day, still regarded as a highly difficult and challenging task, being accountable for the development of many relevant works in the field of systems biology. Its importance in the field is notorious, especially in metabolic engineering, but also in protein engineering problems.

In the presented work we move a step forward in unveiling the molecular determinants for cofactor specificity, using enzyme structural information. Making use of enzyme structural analysis tools and machine learning algorithms we were able to identify interacting couples of cofactor atoms and aminoacid residues in a large enzyme dataset. The proposed findings were successfully applied in the prediction of cofactor specificity of enzymes not structurally characterized, using protein comparative modelling.

To enable high throughput cofactor preference prediction, we developed, trained, implemented and evaluated a method to automatically attribute cofactor specificity preference, when given the aminoacid residue sequence.

We believe that these results represent an important contribution for cofactor engineering problems, and enzyme engineering overall, with minimization of commonly laborious and expensive experimental characterizations. Rational metabolic engineering approaches and strain design endeavors also greatly benefit from these tools with the enhancement of the sensitivity and reliability of metabolic models, through the reduced input of erroneous or redundant reactions, improving the overall performance of metabolic simulations

A webtool was developed to allow a faster and broader reach of the developed work. Through the use of a user friendly query format, researchers without previous skills in enzyme structural engineering can submit their subject enzyme's aminoacid sequence and receive the prediction of its cofactor specificity. Submissions can be performed for one or more aminoacid sequences in the FASTA format at the freely available webserver: http://services.itqb.unl.pt/cofactor-prediction/

3.5 References

[1] W. Tian and J. Skolnick, "How well is enzyme function conserved as a function of pairwise sequence identity?," *J. Mol. Biol.*, vol. 333, no. 4, pp. 863–882, 2003.

[2] D. Lee, O. Redfern, and C. Orengo, "Predicting protein function from sequence and structure," *Nat. Rev. Mol. Cell Biol.*, vol. 8, no. 12, pp. 995–1005, 2007.

[3] K. M. Borgwardt, C. S. Ong, S. Schonauer, S. V. N. Vishwanathan, A. J. Smola, and H.-P. Kriegel, "Protein function prediction via graph kernels," *Bioinformatics*, vol. 21, no. Suppl 1, pp. i47–i56, 2005.

[4] I. V. Tetko, I. V. Rodchenkov, M. C. Walter, T. Rattei, and H. W. Mewes, "Beyond the 'best' match: Machine learning annotation of protein sequences by integration of different sources of information," *Bioinformatics*, vol. 24, no. 5, pp. 621–628, 2008.

[5] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, "Basic local alignment search tool.," *J. Mol. Biol.*, vol. 215, no. 3, pp. 403–10, 1990.

[6] W. R. Pearson and D. J. Lipmant, "Improved tools for biological sequence comparison," *Proc. Natl. Acad. Sci.*, vol. 85, no. April, pp. 2444–2448, 1988.

[7] P. Stothard and D. S. Wishart, “Automated bacterial genome analysis and annotation,” *Curr. Opin. Microbiol.*, vol. 9, no. 5, pp. 505–510, 2006.

[8] A. Bateman, M. J. Martin, C. O’Donovan, M. Magrane, E. Alpi, R. Antunes, B. Bely, M. Bingley, C. Bonilla, R. Britto, B. Bursteinas, H. Bye-AJee, A. Cowley, A. Da Silva, M. De Giorgi, T. Dogan, F. Fazzini, L. G. Castro, L. Figueira, P. Garmiri, G. Georghiou, D. Gonzalez, E. Hatton-Ellis, W. Li, W. Liu, R. Lopez, J. Luo, Y. Lussi, A. MacDougall, A. Nightingale, B. Palka, K. Pichler, D. Poggioli, S. Pundir, L. Pureza, G. Qi, S. Rosanoff, R. Saidi, T. Sawford, A. Shypitsyna, E. Speretta, E. Turner, N. Tyagi, V. Volynkin, T. Wardell, K. Warner, X. Watkins, R. Zaru, H. Zellner, I. Xenarios, L. Bougueleret, A. Bridge, S. Poux, N. Redaschi, L. Aimo, G. ArgoudPuy, A. Auchincloss, K. Axelsen, P. Bansal, D. Baratin, M. C. Blatter, B. Boeckmann, J. Bolleman, E. Boutet, L. Breuza, C. Casal-Casas, E. De Castro, E. Coudert, B. Cuche, M. Doche, D. Dornevil, S. Duvaud, A. Estreicher, L. Famiglietti, M. Feuermann, E. Gasteiger, S. Gehant, V. Gerritsen, A. Gos, N. Gruaz-Gumowski, U. Hinz, C. Hulo, F. Jungo, G. Keller, V. Lara, P. Lemercier, D. Lieberherr, T. Lombardot, X. Martin, P. Masson, A. Morgat, T. Neto, N. Nouspikel, S. Paesano, I. Pedruzzi, S. Pilbout, M. Pozzato, M. Pruess, C. Rivoire, B. Roechert, M. Schneider, C. Sigrist, K. Sonesson, S. Staehli, A. Stutz, S. Sundaram, M. Tognolli, L. Verbregue, A. L. Veuthey, C. H. Wu, C. N. Arighi, L. Arminski, C. Chen, Y. Chen, J. S. Garavelli, H. Huang, K. Laiho, P. McGarvey, D. A. Natale, K. Ross, C. R. Vinayaka, Q. Wang, Y. Wang, L. S. Yeh, and J. Zhang, “UniProt: The universal protein knowledgebase,” *Nucleic Acids Res.*, vol. 45, no. D1, pp. D158–D169, 2017.

[9] S. Placzek, I. Schomburg, A. Chang, L. Jeske, M. Ulbrich, J. Tillack, and D. Schomburg, “BRENDA in 2017: New perspectives and new tools in BRENDA,” *Nucleic Acids Res.*, vol. 45, no. D1, pp. D380–D388, 2017.

[10] Y. Wang, K. Y. San, and G. N. Bennett, “Cofactor engineering for advancing chemical biotechnology,” *Curr. Opin. Biotechnol.*, vol. 24, no. 6, pp. 994–999, 2013.

[11] W. Liu and P. Wang, “Cofactor regeneration for sustainable enzymatic biosynthesis,” *Biotechnol. Adv.*, vol. 25, no. 4, pp. 369–384, 2007.

[12] H. Zhao and W. A. Van Der Donk, “Regeneration of cofactors for use in biocatalysis,” *Curr. Opin. Biotechnol.*, vol. 14, no. 6, pp. 583–589, 2003.

[13] G. A. Khoury, H. Fazelinia, J. W. Chin, R. J. Pantazes, P. C. Cirino, and C. D. Maranas, “Computational design of Candida boidinii xylose reductase for altered cofactor specificity,” *Protein Sci.*, vol. 18, no. 10, pp. 2125–2138, 2009.

[14] J. K. B. Cahn, C. A. Werlang, A. Baumschlager, S. Brinkmann-Chen, S. L. Mayo, and F. H. Arnold, “A General Tool for Engineering the NAD/NADP Cofactor Preference of Oxidoreductases,” *ACS Synth. Biol.*, vol. 6, no. 2, pp. 326–333, 2017.

[15] S. T. Rao and M. G. Rossmann, “Comparison of super-secondary structures in proteins,” *J. Mol. Biol.*, vol. 76, no. 2, pp. 241–256, 1973.

[16] C. Y. Wu, Y. H. Hwa, Y. C. Chen, and C. Lim, “Hidden relationship between conserved residues and locally conserved phosphate-binding structures in NAD(P)-binding proteins,” *J. Phys. Chem. B*, vol. 116, no. 19, pp. 5644–5652, 2012.

[17] Y. H. Hua, C. Y. Wu, K. Sargsyan, and C. Lim, “Sequence-motif Detection of NAD(P)-binding

Proteins: Discovery of a Unique Antibacterial Drug Target," *Sci. Rep.*, vol. 4, pp. 1–7, 2014.

[18] H. M. Geertz-Hansen, N. Blom, A. M. Feist, S. Brunak, and T. N. Petersen, "Cofactory: Sequence-based prediction of cofactor specificity of Rossmann folds," *Proteins Struct. Funct. Bioinforma.*, vol. 82, no. 9, pp. 1819–1828, 2014.

[19] O. Carugo and P. Argos, "NADP-dependent enzymes. I: Conserved stereochemistry of cofactor binding," *Proteins Struct. Funct. Genet.*, vol. 28, no. 1, pp. 10–20, 1997.

[20] J. K. B. Cahn, S. Brinkmann-Chen, T. Spatzal, J. A. Wiig, A. R. Buller, O. Einsle, Y. Hu, M. W. Ribbe, and F. H. Arnold, "Cofactor specificity motifs and the induced fit mechanism in class I ketol-acid reductoisomerases," *Biochem. J.*, vol. 468, no. 3, pp. 475–484, 2015.

[21] W. L. Duax, V. Pletnev, A. Addlagatta, J. Bruenn, and C. M. Weeks, "Rational Proteomics I . Fingerprint Identification and Cofactor Specificity in the Short-Chain Oxidoreductase (SCOR) Enzyme Family," *PROTEINS Struct. Funct. Genet.*, vol. 943, no. February, pp. 931–943, 2003.

[22] S. Brinkmann-Chen, J. K. B. Cahn, and F. H. Arnold, "Uncovering rare NADH-preferring ketol-acid reductoisomerases," *Metab. Eng.*, vol. 26, pp. 17–22, 2014.

[23] E. Di Luccio, R. A. Elling, and D. K. Wilson, "Identification of a novel NADH-specific aldo-keto reductase using sequence and structural homologies," *Biochem. J.*, vol. 400, no. 1, pp. 105–114, 2006.

[24] D. Cui, L. Zhang, S. Jiang, Z. Yao, B. Gao, J. Lin, Y. A. Yuan, and D. Wei, "A computational strategy for altering an enzyme in its cofactor preference to NAD(H) and/or NADP(H)," *FEBS J.*, vol. 282, no. 12, pp. 2339–2351, 2015.

[25] C. C. Marohnic, M. C. Bewley, and M. J. Barber, "Engineering and characterization of a NADPH-utilizing cytochrome b 5 reductase," *Biochemistry*, vol. 42, no. 38, pp. 11170–11182, 2003.

[26] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, "The protein data bank.," *Nucleic Acids Res.*, vol. 28, no. 1, pp. 235–242, 2000.

[27] P. Larrañaga, B. Calvo, R. Santana, C. Bielza, J. Galdiano, I. Inza, J. A. Lozano, R. Armañanzas, G. Santafé, A. Pérez, and V. Robles, "Machine learning in bioinformatics," *Brief. Bioinform.*, vol. 7, no. 1, pp. 86–112, 2006.

[28] N. Nagamine and Y. Sakakibara, "Statistical prediction of protein - Chemical interactions based on chemical structure and mass spectrometry data," *Bioinformatics*, vol. 23, no. 15, pp. 2004–2012, 2007.

[29] L. Jacob and J. P. Vert, "Protein-ligand interaction prediction: An improved chemogenomics approach," *Bioinformatics*, vol. 24, no. 19, pp. 2149–2156, 2008.

[30] H. Yu, J. Chen, X. Xu, Y. Li, H. Zhao, Y. Fang, X. Li, W. Zhou, W. Wang, and Y. Wang, "A systematic prediction of multiple drug-target interactions from chemical, genomic, and pharmacological data," *PLoS One*, vol. 7, no. 5, 2012.

[31] Z. He, J. Zhang, X.-H. Shi, L.-L. Hu, X. Kong, Y.-D. Cai, and K.-C. Chou, "Predicting Drug-Target Interaction Networks Based on Functional Groups and Biological Features," *PLoS One*, vol. 5, no. 3, p. e9603, 2010.

[32] P. Korkuć and D. Walther, "Physicochemical characteristics of structurally determined metabolite-protein and drug-protein binding events with respect to binding specificity," *Front. Mol. Biosci.*, vol. 2, no. September, pp. 1–20, 2015.

[33] T. Hamelryck and B. Manderick, "PDB file parser and structure class implemented in Python," *Bioinformatics*, vol. 19, no. 17, pp. 2308–2310, 2003.

[34] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and É. Duchesnay, "Scikit-learn: Machine Learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2012.

[35] S. Van Der Walt, S. Africa, and M. S. Feb, "The NumPy array : a structure for efficient numerical computation," pp. 1–8, 2011.

[36] B. W. Matthews, "COMPARISON OF THE PREDICTED AND OBSERVED SECONDARY STRUCTURE OF T4 PHAGE LYSOZYME," *Biochim. Biophys. Acta*, vol. 405, pp. 442–451, 1975.

[37] J. Swets, "Measuring the accuracy of diagnostic systems," *Science (80-.).*, vol. 240, no. 4857, pp. 1285–1293, 1988.

[38] B. Webb and A. Sali, *Comparative protein structure modeling using MODELLER*, vol. 2014. 2014.

[39] T. Smith and M. Waterman, "Identification of common molecular subsequences," *J. Mol. Biol.*, vol. 147, no. 3, pp. 195–197, 1981.

[40] W. R. Pearson, "Empirical Statistical Estimates for Sequence Similarity Searches," *J. Mol. Biol.*, vol. 276, pp. 71–84, 1998.

[41] D. Box, D. Ehnebuske, G. Kakivaya, A. Layman, N. Mendelsohn, H. Nielsen, S. Thatte, and D. Winer, "Simple Object Access Protocol (SOAP) 1 . 1," *W3C - World Wide Web Consort. Note*, no. October, 2000.

[42] C. Cortes and V. Vapnik, "Support Vector Networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, 1995.

[43] S. Watanabe, M. Yamada, I. Ohtsu, and K. Makino, "α-Ketoglutaric semialdehyde dehydrogenase isozymes involved in metabolic pathways of D-glucarate, D-galactarate, and hydroxy-L-proline: Molecular and metabolic convergent evolution," *J. Biol. Chem.*, vol. 282, no. 9, pp. 6685–6695, 2007.

[44] X. Liu, Y. Dong, J. Zhang, A. Zhang, L. Wang, and L. Feng, "Two novel metal-independent long-chain alkyl alcohol dehydrogenases from Geobacillus thermodenitrificans NG80-2," *Microbiology*, vol. 155, pp. 2078–2085, 2009.

[45] B. Ahvazi, R. Coulombe, M. Delarge, M. Vedadi, L. Zhang, E. Meighen, and A. Vrielink,

"Crystal structure of the NADP+-depe,indent aldehyde dehydrogenase from *Vibrio harveyi*: structural implications for cofactor specificity and affinity.," *Biochem. J.*, vol. 349 Pt 3, pp. 853–61, 2000.

[46] S. Leitgeb, B. Petschacher, D. K. Wilson, and B. Nidetzky, "Fine tuning of coenzyme specificity in family 2 aldo-keto reductases revealed by crystal structures of the Lys-274 ??? Arg mutant of Candida tenuis xylose reductase (AKR2B5) bound to NAD+ and NADP+," *FEBS Lett.*, vol. 579, no. 3, pp. 763–767, 2005.

[47] B. PETSCHACHER, S. LEITGEB, K. L. KAVANAGH, D. K. WILSON, and B. NIDETZKY, "The coenzyme specificity of Candida tenuis xylose reductase (AKR2B5) explored by site-directed mutagenesis and X-ray crystallography," *Biochem. J.*, vol. 385, no. 1, pp. 75–83, 2005.

[48] S. Banta, B. A. Swanson, S. Wu, A. Jarnagin, and S. Anderson, "Alteration of the specificity of the cofactor-binding pocket of Corynebacterium 2,5-diketo-D-gluconic acid reductase A," *Protein Eng.*, vol. 15, no. 2, pp. 131–140, 2002.

[49] S. Banta, B. A. Swanson, S. Wu, A. Jarnagin, and S. Anderson, "Optimizing an artificial metabolic pathway: Engineering the cofactor specificity of Corynebacterium 2,5-diketo-D-gluconic acid reductase for use in vitamin C biosynthesis," *Biochemistry*, vol. 41, no. 20, pp. 6226–6236, 2002.

[50] M. J. Rane and K. C. Calvo, "Reversal of the nucleotide specificity of ketol acid reductoisomerase by site-directed mutagenesis identifies the NADPH binding site," *Arch. Biochem. Biophys.*, vol. 338, no. 1, pp. 83–89, 1997.

[51] N. Shiraishi, C. Croy, J. Kaur, and W. H. Campbell, "Engineering of Pyridine Nucleotide Specificity of Nitrate Reductase: Mutagenesis of Recombinant CytochromebReductase Fragment ofNeurospora crassaNADPH:Nitrate Reductase," *Arch. Biochem. Biophys.*, vol. 358, no. 1, pp. 104–115, 1998.

[52] M. Kostrzynska, C. R. Sopher, and H. Lee, "Mutational analysis of the role of the conserved lysine-270 in the _Pichia stipitis_ xylose reductase.," *FEMS Microbiol.Lett.*, vol. 159, no. March, pp. 107–112, 1998.

[53] L. Liang, J. Zhang, and Z. Lin, "Altering coenzyme specificity of Pichia stipitis xylose reductase by the semi-rational approach CASTing," *Microb. Cell Fact.*, vol. 6, pp. 1–11, 2007.

[54] M. H. M. Eppink, K. M. Overkamp, H. A. Schreuder, and W. J. H. Van Berkel, "Switch of coenzyme specificity of p-hydroxybenzoate hydroxylase," *J. Mol. Biol.*, vol. 292, no. 1, pp. 87–96, 1999.

[55] C. L. Elmore and T. D. Porter, "Modification of the nucleotide cofactor-binding site of cytochrome P-450 reductase to enhance turnover with NADH in vivo," *J. Biol. Chem.*, vol. 277, no. 50, pp. 48960–48964, 2002.

[56] T. R. Dambe, A. M. Kühn, T. Brossette, F. Giffhorn, and A. J. Scheidig, "Crystal structure of NADP(H)-dependent 1,5-anhydro-D-fructose reductase from Sinorhizobium morelense at 2.2 Å resolution: Construction of a NADH-accepting mutant and its application in rare sugar synthesis," *Biochemistry*, vol. 45, no. 33, pp. 10030–10042, 2006.

[57] R. Chen, a Greer, and a M. Dean, "A highly active decarboxylating dehydrogenase with

rationally inverted coenzyme specificity.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 92, no. 25, pp. 11666–11670, 1995.

[58] M. J. Barber, B. A. Notton, C. J. Kay, and L. P. Solomonson, "Chloride Inhibition of Spinach Nitrate Reductasel," pp. 70–74, 1989.

[59] A. H. Ehrensberger, R. A. Elling, and D. K. Wilson, "Structure-guided engineering of xylitol dehydrogenase cosubstrate specificity," *Structure*, vol. 14, no. 3, pp. 567–575, 2006.

[60] S. Watanabe, T. Kodaki, and K. Makino, "Complete reversal of coenzyme specificity of xylitol dehydrogenase and increase of thermostability by the introduction of structural zinc," *J. Biol. Chem.*, vol. 280, no. 11, pp. 10340–10349, 2005.

[61] R. Chen, A. F. Greer, A. M. Dean, and J. H. Hurley, "Redesigning secondary structure to invert coenzyme specificity in isopropylmalate dehydrogenase," *Proc. Natl. Acad. Sci.*, vol. 93, pp. 12171–12176, 1996.

[62] a Galkin, L. Kulakova, T. Ohshima, N. Esaki, and K. Soda, "Construction of a new leucine dehydrogenase with preferred specificity for NADP+ by site-directed mutagenesis of the strictly NAD+-specific enzyme.," *Protein Eng.*, vol. 10, no. 6, pp. 687–690, 1997.

[63] A. Rodríguez-Arnedo, M. Camacho, F. Llorca, and M. J. Bonete, "Complete reversal of coenzyme specificity of isocitrate dehydrogenase from Haloferax volcanii," *Protein J.*, vol. 24, no. 5, pp. 259–266, 2005.

CHAPTER 4

Applying NiCofactor to the curation of NAD(P)(H) usage in Genome-scale metabolic models

The information presented in this Chapter is being prepared for submission to a peer reviewed journal:

Resende T., Soares C., Rocha I. Analysis and curation of NAD(P)(H) usage in Genome-scale metabolic models

4.1 Introduction

Previously to the genetic revolution, strain design and improvement was performed naturally, with selective breeding and environmental pressure, or by the implementation of random genetic mutagenesis followed by positive phenotype selection. With the evolution of genetic and metabolic engineering, metabolism is now better understood and controlled, with more rational and precise approaches being implemented [1]–[5]. Currently, despite the vast knowledge on metabolism, issues still arise when designing engineered strains, as the complexity of metabolic pathways divert the metabolism, causing undesired phenotypes with unclear causes.

Advances in genome sequencing allowed the number of sequenced organisms to increase dramatically in the last decade. In this post-genomic era, functional annotation became a main concern of bioinformatics and systems biology [6], [7]. Systems biology aims to comprehend the relations between genes, proteins and metabolites of an organism's metabolism through the computational modelling of these complex biological systems [8]. With the ever increasing amount of genomic functional information available, the reconstructions and simulation of genome-scale metabolic models (GEMs) is a prolific methodology to achieve such purpose.

GEMs comprehend information on the complex network of biochemical reactions occurring inside an organism in the form of stoichiometric equations, representing a detailed depiction of the organism's metabolism. GEMs can predict the phenotype of the organism according to the set environmental conditions, allowing for a detailed analysis of the internal metabolite fluxes, as well as the genotype-phenotype associations. This methodology has been successfully implemented in multiple organisms, representing prokaryotic, eukaryotic and also archaea species [9].

Along with wild-type phenotype, gene knockout or overexpression is also feasible, modelling the redirection of fluxes trough the existing pathways for the optimization of specific compound production [10]–[16]. The correct representation of the metabolic reactions occurring in the metabolism of an organism is crucial for an accurate simulation of such organisms’ metabolism, making the process of GEM reconstructions a laborious and precise task, involving multiple curation steps. This process has been discussed and described by several authors, along with the development of several tools with the intent of semi-automatizing GEM reconstruction [9], [17]–[23]. Once reconstructed, GEMs can be simulated using multiple different approaches and algorithms, being Flux Balanced Analysis (FBA) [24] and parsimonious Flux Balanced Analysis (pFBA) [25] two of the most used strategies. These methods use linear programming to maximize an objective function, usually Biomass growth, through the distribution of fluxes in the model.

Despite all information available and developed methodologies, GEM reconstruction is still prone to the insertion of errors, forming misleading reactions and model inconsistencies that can render this effort ineffective. Experimental determination of a protein function is expensive and time-consuming, making it unfeasible for the growing amount of sequences [26]–[28]. By largely relying on function annotation from sequence homology, errors and inconsistencies spread easily when this process is not done carefully due to overly unconstrained homology search metrics [6], [29]. Also, despite all the important information stored in a gene, some characteristics, such us cofactor usage of the gene encoded reactions are very difficult to retrieve by gene sequence homology comparison.

Cofactors are key elements in the overall metabolism regulation of an organism. Every pathway encompasses reactions that make use of different cofactors to allow the successful biochemical transitions within an enzyme. NAD(H) and NADP(H) are among the most used cofactors in cell metabolism, being the precise description of their availability utterly important for the correct simulation of catabolic and anabolic processes [30].

Despite a certain degree of promiscuity associated to enzymes using NAD(P)(H) as cofactors, the structure of an enzyme always fits better with one of them [31], [32]. Also, cofactor promiscuity representation in GEMs is a very dangerous procedure, once the natural representation of this phenomenon is to insert duplicated reactions, changing only the cofactor, which leads to an input of uncertain reactions, as well as the creation of possible futile cycles, where the regeneration of one of the cofactors occurs artificially.

In this work we perform the curation of genome-scale metabolic models through the correct characterization of reactions using the cofactors NAD(P)(H). For that, we apply our in house developed software NiCofactor, which uses enzyme structural information and machine learning, to 59 different reconstructed GEMs belonging to 47 different strains. Results help depicting the state of cofactor curation in GEM reconstruction, as well as the importance of accurate cofactor specificity attribution. Furthermore, the correction of the most recent model from *S. cerevisiae* is used in simulations and the resulting fluxes compared to the original and *in vivo* flux estimations for a performance assessment.

4.2 Methods

4.2.1 GEM download and Aminoacid sequence retrieval

Aminoacid sequences related to reactions using NAD(P)(H) were retrieved from an *in-house* developed framework used to integrate and curate all compound present in several models available in the literature. NAD(H) and NADP(H) compounds searched and the respective reactions retrieved. From the retrieved reactions, associated genes and aminoacid sequences were collected.

To perform reaction corrections and model simulations, Yeast 7.6 model [33] was downloaded from the project's website: http://sourceforge.net/projects/yeast/files/.

4.2.2 NiCofactor aminoacid sequence processing

NAD(P)(H) cofactor specificity prediction for each of the genes encoding reactions in all analyzed models using NAD(P)(H) was performed using NiCofactor. Genes were retrieved in their protein aminoacid sequence organized in the Fasta format prior to processing. Fasta files, containing the subject protein sequences were processed automatically in the software NiCofactor and the prediction results were outputted in text format containing the predicted cofactors, as well as the confidence score for each sequence.

4.2.3 Simulations with GEMs

GEM Yeast 7.6 was simulated using OptFlux 3.2.8 [34] set with the following *in silico* environmental conditions: glucose with an uptake rate of 1.15 mmol/gCDW·h, ammonia with unconstrained uptake, sulfate with unconstrained uptake, phosphate with unconstrained uptake and oxygen with unconstrained uptake. Parsimonius Flux Balanced Analysis (pFBA) [25] was used for the calculation of internal flux distributions and the set objective function was the maximization of biomass growth. Modifications in the reactions of the model were directly made in the original SBML file [35].

4.3 Results

4.3.1 Characterization of the GEM dataset

The analyzed GEMs represent and model the metabolism of multiple organisms, from different taxa. This dataset encompasses all GEMs published between 2005 and 2016 for which a functional model file could be retrieved. Some models for the same organism, as for example *S. cerevisiae* models, were developed using previously existing models, and represent an updated version of the organism metabolism representations. This is due to the ongoing experimental validation of multiple parameters and also the ever increasing amount of available information in public databases.

Many other organisms are represented in the returned GEMs. Table 4.1 depicts all organisms represented, as well as their correspondent GEM, year of GEM publication and taxonomic domain.

Table 4.1 – Complete list of the retrieved and analyzed GEM, encompassing Model name, year of publication, organism name, taxa and reference.

Model	Year	Organism	Taxa	Ref
*i*AZ900	2010	*Saccharomyces cerevisiae* S288c	Eukaryota	[36]
*i*IN800	2008	*Saccharomyces cerevisiae* S288c	Eukaryota	[37]
*i*JO1366	2011	*Escherichia coli* str. K-12 substr. MG1655	Bacteria	[38]
*i*MM904	2009	*Saccharomyces cerevisiae* S288c	Eukaryota	[39]
*i*NJ661m	2010	*Mycobacterium tuberculosis* H37Rv	Bacteria	[40]
*i*TO977	2013	*Saccharomyces cerevisiae* S288c	Eukaryota	[41]
Yeast 7.6	2015	*Saccharomyces cerevisiae* S288c	Eukaryota	[33]
Yeast 6	2013	*Saccharomyces cerevisiae* S288c	Eukaryota	[42]
*i*Abaylyiv4	2008	*Acinetobacter* sp. ADP1	Bacteria	[43]
*i*AI558	2015	*Moorella thermoacetica* ATCC 39073	Bacteria	[44]
*i*BT721	2012	*Lactobacillus plantarum* WCFS1	Bacteria	[45]
*i*Cce806	2012	*Cyanothece* sp. ATCC 51142	Bacteria	[46]
*i*CG230	2012	*Blattabacterium* sp. str. BPLAN	Bacteria	[47]
*i*CM925	2011	*Clostridium beijerinckii* NCIMB 8052	Bacteria	[48]
*i*CR744	2009	*Rhodoferax ferrireducens* T118	Bacteria	[49]
*i*Cyc792	2013	*Cyanothece* sp. PCC 7424	Bacteria	[50]
*i*Cyj826	2013	*Cyanothece* sp. PCC 7822	Bacteria	[50]
*i*Cyn731	2013	*Cyanothece* sp. PCC 7425	Bacteria	[50]
*i*Cyp752	2013	*Cyanothece* sp. PCC 8801	Bacteria	[50]
*i*JB785	2016	*Synechococcus elongatus* PCC 7942	Bacteria	[51]
*i*JL432	2008	*Clostridium acetobutylicum* ATCC 824	Bacteria	[52]
*i*JL480	2016	*Streptococcus pyogenes* NZ131	Bacteria	[53]
*i*JS747	2009	*Geobacter metallireducens* GS-15	Bacteria	[54]
*i*MF721	2014	*Pseudoalteromonas haloplanktis* TAC125	Bacteria	[55]
*i*MG746	2013	*Methanosarcina barkeri* str. Fusaro	Archaea	[56]
*i*MP240	2013	*Blattabacterium* sp. (Blattella germanica) str. Bge	Bacteria	[57]
*i*NF518	2013	*Lactococcus lactis* subsp. *cremoris* MG1363	Bacteria	[58]
*i*PS189	2009	*Mycoplasma genitalium* G37	Bacteria	[59]

*i*RR1083	2009	*Salmonella enterica* subsp. *serovar* Typhimurium	Bacteria	[60]
*i*WZ663	2012	*Ketogulonicigenium vulgare* WSH-001	Bacteria	[61]
*i*TM560	2011	*Neisseria meningitidis* MC58	Bacteria	[62]
*i*CAC490	2012	*Clostridium acetobutylicum* ATCC 824	Bacteria	[63]
*i*Cac802	2014	*Clostridium acetobutylicum* ATCC 824	Bacteria	[64]
*i*Cyt773	2012	*Cyanothece* sp. ATCC 51142	Bacteria	[65]
*i*EM439	2016	*Zymomonas mobilis* subsp. *mobilis* ATCC 10988	Bacteria	[66]
*i*JH728	2016	*Synechococcus* sp. PCC 7002	Bacteria	[67]
*i*JP815	2008	*Pseudomonas putida* KT2440	Bacteria	[68]
*i*JP962	2011	*Pseudomonas putida* KT2440	Bacteria	[69]
*i*PB890	2015	*Pseudomonas stutzeri* A1501	Bacteria	[70]
*i*SO783	2010	*Shewanella oneidensis* MR-1	Bacteria	[71]
*i*CG238	2012	*Blattabacterium* sp. (Blattella germanica) str. Bge	Bacteria	[47]
*i*IB700	2005	*Streptomyces coelicolor* A3(2)	Bacteria	[72]
*i*MK1208	2014	*Streptomyces coelicolor* A3(2)	Bacteria	[73]
*i*MO1056	2008	*Pseudomonas aeruginosa* PAO1	Bacteria	[74]
*i*MP429	2009	*Streptococcus thermophilus* LMG 18311	Bacteria	[75]
*i*MZ1055	2013	*Bacillus megaterium* WSH-002	Bacteria	[76]
*i*NV706	2014	*Enterococcus faecalis* V583	Bacteria	[77]
*i*RM588	2006	*Geobacter sulfurreducens* PCA	Bacteria	[78]
*i*Rsp1095	2011	*Rhodobacter sphaeroides* 2.4.1	Bacteria	[79]
*i*Rsp1140	2013	*Rhodobacter sphaeroides* 2.4.1	Bacteria	[80]
*i*JL846	2014	*Lactobacillus casei* LC2W	Bacteria	[81]
*i*JSPpropionicus	2011	*Pelobacter propionicus* DSM 2379	Bacteria	[82]
*i*YLW1028	2015	*Actinoplanes* sp. SE50/110	Bacteria	[83]
*i*JSPcarbinolicus	2011	*Pelobacter carbinolicus* DSM 2380	Bacteria	[82]
*i*KY620	2015	*Arthrospira platensis* NIES-39	Bacteria	[84]
*i*AM388	2011	*Campylobacter jejuni* subsp. *jejuni* NCTC 11168	Bacteria	[85]
*i*MR539	2016	*Methanococcus maripaludis* S2	Archaea	[86]
*i*Cyh755	2013	*Cyanothece* sp. PCC 8802	Bacteria	[50]
*i*YS432	2009	*Corynebacterium glutamicum* ATCC 13032	Bacteria	[87]

Although 59 GEMs were retrieved, only 47 organisms are represented, being 44 bacteria, 1 eukaryota (*S. cerevisiae*) and 2 archaea (*Methanosarcina barkeri* str. Fusaro and *Methanococcus maripaludis* S2). Organisms with multiple GEMs are *S. cerevisiae*, with 6; *Clostridium acetobutylicum* ATCC 824, with three (*i*Cac802, *i*CAC490 and *i*JL432); *Cyanothece* sp. ATCC 51142, with two

(*i*Cyt773 and *i*Cce806); *Pseudomonas putida* KT2440, with two (*i*JP962 and *i*JP815) *Blattabacterium* sp. (Blattella germanica) str. Bge, with two (*i*MP240 and *i*CG238); *Streptomyces coelicolor* A3, with two (*i*IB700 and *i*MK1208); *Rhodobacter sphaeroides* 2.4.1 with two (*i*Rsp1140 and *i*Rsp1095).

Despite the fact that reactions present in the GEM represent the metabolism encoded by the genome, GEMs do not compulsorily need gene information associated to their reactions, being often incorporated reactions that are not associated to any GPR in the process of gap-filling [88], [57], [89]. Albeit this fact, such reactions must have a lower level of confidence, and be treated as such during model simulation.

Figure 4.1 depicts in a graph the total amount of reactions and genes and the number of reactions and genes associated with NAD(P)(H) and their ratio.

The model with the highest amount of incorporated genes is *E. coli*'s *i*JO1366 with 1367 genes, being the amount of reactions in the model 2253. The fact that this bacterium is one of the most studied and industrially better suited organisms contributes for this high representation of genes [90]. The average amount of genes and reactions present in the analyzed models is 743 and 1005, respectively, while the average amount of genes related to NAD(P)(H) is 112 and the average amount of NAD(P)(H) related reactions, 122.

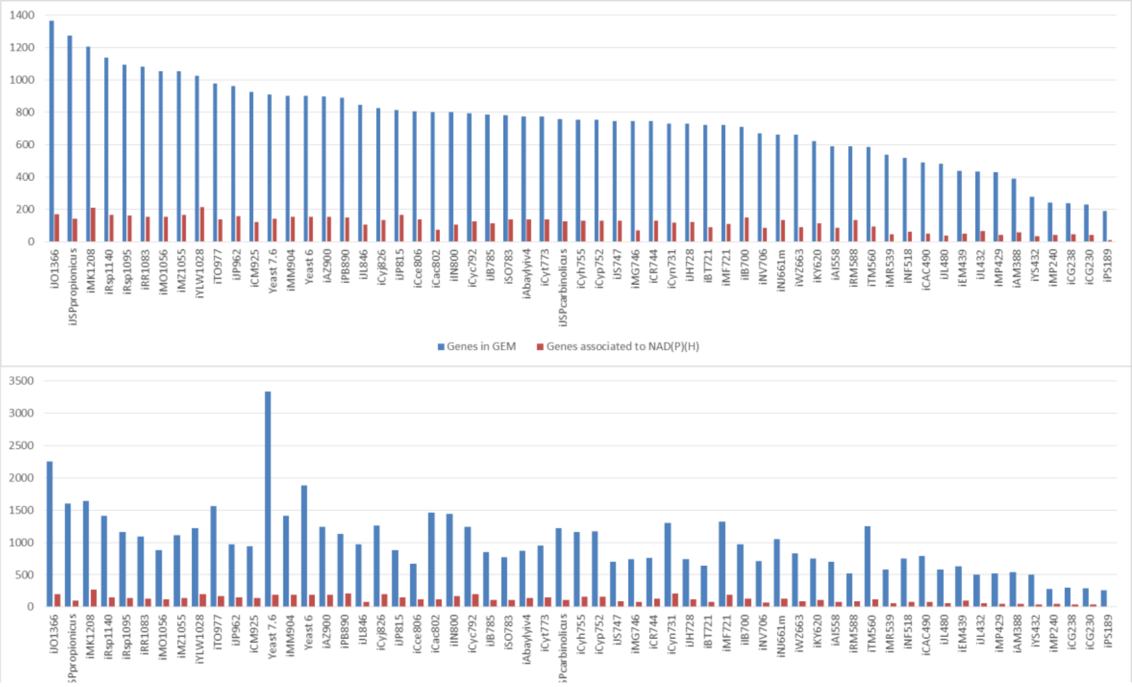


Figure 4.1 – 1) Comparison between the total amount of genes included in each GEM (blue) with the number of genes in each GEM associated to reactions using NAD(P)(H) (red). 2) Comparison between the total amount of reactions included in each GEM (blue) with the number of reactions using NAD(P)(H) (red).

This analysis shows that, overall, the percentage of NAD(P)(H) related genes incorporated in the models attend to 15.1% of the total gene count, while reactions related to NAD(P)(H) account for 12.1% of the total amount of reactions. Given the fact that GEM reconstruction integrates a comprehensive number of reactions present in the organism, through multiple pathways, capturing several aspects of its metabolism, such high percentage of genes and reactions associated with NAD(P)(H) reinforce the important contribution of these cofactors in the overall metabolism of organisms.

Despite being the model with the highest amount of genes, *i*JO1366 does not have the highest amount of reactions. Yeast 7.6 has 3334 included reactions, while having only 910 included genes. The model with the highest amount of genes related to NAD(P)(H), is *i*YLW1028, a GEM from *Actinoplanes* sp. SE50/110, with 213 (20.7%) of the total 1028 genes related to NAD(P)(H), while it has 1219 reactions, being 196 (16.1%) of which associated to NAD(P)(H).

The model with the highest amount of NAD(P)(H) associated reactions is *i*MK1208, a GEM for *Streptomyces coelicolor*, with 269 (16.4%), from a total of 1643, reactions related to NAD(P)(H).

The model with the lowest amount (and percentage) of reactions and genes is *i*PS189, from *Mycoplasma genitalium* G37, with only 189 included genes and 262 reactions, being 9 (4.8%) genes and 10 (3.8%) reactions related to NAD(P)(H).

Despite not being the model with the highest amount of NAD(P)(H) associated reactions, model *i*RM588 from *Geobacter sulfurreducens* PCA has the highest percentage, in the analyzed GEMs, of NAD(P)(H) reactions, with 22.7% of NAD(P)(H) related reactions. In the case of the genes, *i*PB890 from *Pseudomonas stutzeri* A1501 has the highest percentage of NAD(P)(H) related genes, with 18.6% of NAD(P)(H) genes.

4.3.2 Model NAD(P)(H) cofactor usage analysis

Despite being an updated version, or a complete new version, depending on the author, GEMs from the same organism naturally tend to share most of their reactions and GPRs. In order to prevent de duplications of results, and the incorrect representation of the performed GEM analysis, using NiCofactor, only the most recent model versions of each organism was analyzed. This reduced the amount of GEMs from 59 to 47, being 12 GEMs removed.

Being this analysis specific for NAD(P)(H) related reactions, only reactions containing NAD(P)(H) and a GPR association were analyzed. Biomass and NAD(P) transhydrogenase (THD) reactions, accounting for 110 reactions across all 47 models, were also discarded, as the NAD(P)(H) specificity problem does not apply here.

The final dataset was composed by 5081 different genes associated to 5472 reactions from the 47 GEMs.

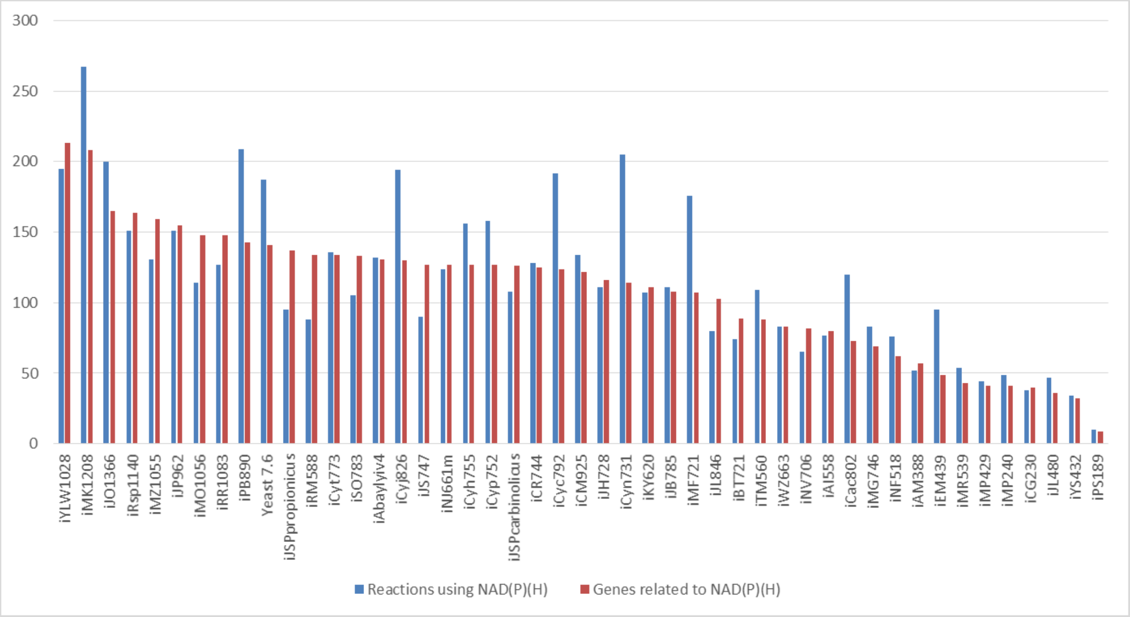


Figure 4.2 – Total amount of reactions using NAD(P)(H) (blue) and genes associated to reactions using NAD(P)(H) (red) in the selected 47 GEM, excluding the biomass equation and THD reactions (when occurring).

Figure 4.2 depicts the number of genes and reactions associated to NAD(P)(H) in the analyzed 47 GEMs. With the reduction in the amount of analyzed models, due to the exclusion of models from the same organism, and also the exclusion of biomass and THD reactions, the above displayed data contains the number of genes and reactions that, in each model, are linked to the utilization of one of the cofactors NAD(P)(H). When analyzing the exposed data, one can observe that there is a great variation in the number of reactions and genes associated to NAD(P)(H) dispersed through the analyzed models. However, such differences are attenuated when considering the number of genes and reactions included in the respective models. Such variation in total amount of gene and reaction incorporation is linked to several aspects of GEM reconstruction, being the total genome size one of the main contributors. GEM author curation methodology, available genomic information or experimental data also play a crucial role in the overall representation of the organism’s metabolism by the GEM [17], [18]. Given the set of genes and reactions displayed above, it is observable that the GEM encompassing more reactions is *i*KM1208 with 267, while *i*YLW1028 is the GEM with most genes included, with 213.

A single gene is often responsible for the transcription and subsequent translation of an enzyme capable of catabolizing multiple different substrates into several products. It is also truth that, regularly, several genes in an organism's genome encode for the same enzyme [91]. In order to correctly reconstruct a GEM, transcribing, *in silico*, a real depiction of the reality, reactions are associated to Gene-Protein-Reactions (GPRs) that depict events such as enzyme complexes, multiple reactions encoded by the same gene or multiple genes associated to one reaction only [18].

In the presented models, the NAD(P)(H) related gene associated with most reactions is the gene Sco1814, from GEM *i*MK1208, associated to 92 different reactions in the model. This gene encodes for an Enoyl-[acyl-carrier-protein] reductase, part of the fatty acid biosynthesis.

With respect to GPRs, the biggest GPR present in the models belong to *i*JSPpropionicus, a GEM from *Pelobacter propionicus* DSM 2379, composed by a total of 50 different genes. The encoded reaction represents the transformation of menaquinone 7 into menaquinol 7 by NADH Dehydrogenase. Notwithstanding the large size of the biggest GPR present in the models, the average GPR size corresponds to combinations of 2 genes.

With respect to the cofactor used in the presented reactions, the amount of reactions using NAD(H) is 2729, while NADP(H) is used in 2743 reactions. As for gene-cofactor association in models, from the total 5081 genes present in the model's reactions GPRs, 2539 genes are present in GPRs from reactions using NAD(H), while 1975 genes are in GPRs from reactions using NADP(H). There are also 567 genes that compose GPRs linked to different reactions using both NAD(H) or NADP(H), which enhances the importance of this study for a better GEM curation.

4.3.3 NiCofactor cofactor prediction results

The developed software NiCofactor, which predicts NAD(P)(H) cofactor specificity, was applied in the analysis of the genes associated to NAD(P)(H) using reactions, present in the retrieved reactions dataset from the gathered 47 GEMs. In the first step, NiCofactor works by searching, for a given gene, protein sequence homologues having their tridimensional structure characterized. Only structures from NAD(P)(H) using enzymes, and bound to one of the cofactors, are allowed as structural templates for performing the structure modeling of the target gene.

Figure 4.3 depicts the results from the analyzed genes. From the total 5081 different genes encoding reactions using NAD(P)(H) in the 47 retrieved GEM, 3232 were found to have a suitable structure template to perform structure comparative modeling, and having their cofactor specificity prediction performed. Once the tridimensional structure of a protein is modelled, NiCofactor makes use of a machine learning algorithm to perform a prediction on its cofactor specificity, outputting also a confidence score. For exceptional prediction accuracy (>90%) from NiCofactor, a score threshold is applied, being only accepted predictions with a prediction score equal to or greater than 0.8, from now on called predicted genes.

The total amount of predicted genes is 2334, representing 72.2% of the genes with a suitable structural template. From these, 1659 genes are associated with reactions using the matching predicted cofactor, while 436 correspond to mismatches. The remaining 239 genes are associated to reactions using both NAD(H) and NADP(H).

The 2334 predicted genes are present in GPRs encoding for 3763 metabolic reactions. From these, 2621 reactions (69.7%) match the predicted cofactor of the genes represented in their GPR, while 1142 (30.3%) have in their GPR genes that are predicted as using a different cofactor.

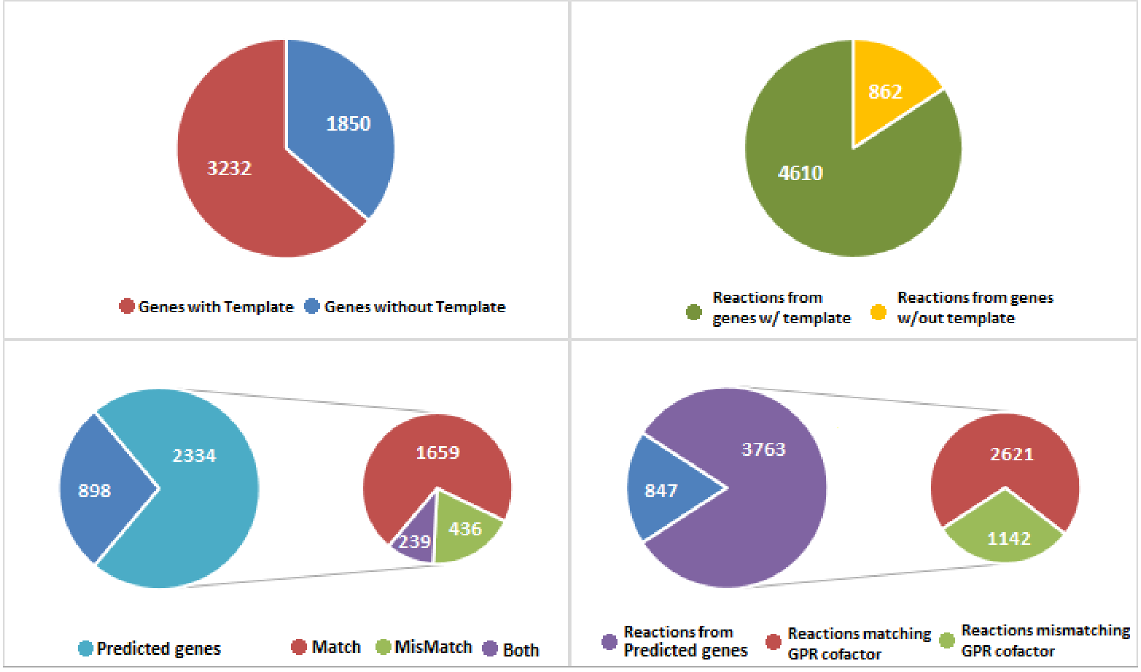


Figure 4.3 - Left top: Amount of Genes found to have a suitable structural template. In red are displayed the genes that have their structure characterized, or have a suitable structure template. In blue, the genes that do not have structural template. Left Bottom: From the total amount of genes with template, the amount of genes that had a prediction score equal or above 0.8 are called predicted genes. From the total amount of predicted genes, the amount that match with encoded reactions, the mismatches and the genes that encode reactions with both cofactors. Right top: reactions using NAD(P)(H) as cofactors that have genes with template in their GPR. Green is the amount of reactions whose genes have structure template. Right bottom: in purple is amount of reactions encoded by the predicted genes. In red are the reactions that match the gene cofactor and in green those that do not.

The below displayed table 4.2 shows the distribution of the genes and reactions in the analyzed set, across all models.

Table 4.2 – GEM distribution of the amount of genes with template and predicted genes, along with the reactions from genes with template and the reactions from the predicted genes.

Model	Genes with template	Predicted genes	Reactions from genes w/ template	Reactions from Predicted genes
iJO1366	108	82	176	152
iNJ661m	83	61	113	100
Yeast 7.6	98	77	148	126
iAbaylyiv4	93	73	124	107
iAl558	54	39	70	56
iBT721	79	54	69	46
iCG230	29	20	37	28
iCM925	94	54	119	81
iCR744	78	57	104	84
iCyc792	77	60	154	136

iCyj826	82	53	155	126
iCyn731	67	46	159	124
iCyp752	66	46	116	88
iJB785	52	35	79	59
iJL480	31	23	41	32
iJS747	65	48	78	65
iMF721	85	65	159	141
iMG746	48	31	65	49
iMP240	28	21	47	28
iNF518	55	40	73	58
iPS189	7	7	9	9
iRR1083	102	79	115	97
iWZ663	52	39	72	59
iTM560	50	38	98	82
iCac802	60	40	107	81
iCyt773	66	45	96	71
iEM439	37	24	82	67
iJH728	54	40	77	59
iJP962	109	80	138	116
iPB890	99	79	186	174
iSO783	75	54	97	83
iMK1208	126	91	248	232
iMO1056	106	79	104	82
iMP429	34	26	42	35
iMZ1055	127	101	119	106
iNV706	67	39	63	43
iRM588	65	46	81	67
iRsp1140	106	79	131	104
iJL846	83	53	77	49
iJSPpropionicus	9	4	8	5
iYLW1028	133	97	165	147
iJSPcarbinolicus	82	62	93	71
iKY620	53	38	86	69
iAM388	32	23	42	33
iMR539	34	26	46	37
iCyh755	65	43	113	81
iYS432	27	17	29	18

In the presented table are displayed, for each model individually, the number of genes from the model that have a suitable structure template, as well as the number of predicted genes. The average amount of genes with a template for each model, taking into account the total amount of genes in the model is 66.08%, while the average percentage of predicted genes, from the pool of genes in each model, is 47%, which, despite appearing to be a small percentage, when compared with the percentage of genes that have a template (66.08%), reveals itself as a good result. The model with the lowest percentage of predicted genes is *i*JSPpropionicus, with only 2.92% of the genes in the model having a cofactor specificity prediction. The GEM with highest amount of genes predicted in the model is *i*PS189, with 77.8%.

When thoroughly analyzing the reactions encoded by the analyzed genes, present in their correspondent GPRs, it is possible to observe that the overall reach of the predictions performed goes much further, having a substantially higher percentage of reactions with their cofactor usage scrutinized. This is due to the fact that several genes are present in the GPR of multiple reactions, enlarging this way the percentage of reactions analyzed.

When analyzing the above displayed results we can observe that an average of 67.9% of the reactions present in each model have predicted genes in their GPR. A total of 76% of NAD(P)(H) reactions present in *E. coli*'s *i*JO1366 have, in their GPR, predicted genes, while 67.4% had the same for *S. cerevisiae*'s Yeast 7.6.

These results show that the developed method is able to reach a large portion of the genes and reactions present in several GEMs, being able to analyze the cofactor specificity of several genes and the cofactor usage of a significant amount of reactions.

In table 4.3 it is possible to observe the amount of reactions matching cofactor usage with gene cofactor prediction.

Table 4.3 – GEM distribution of the amount of predicted genes matching and mismatching reaction cofactor usage, as well as the amount of genes that encode reactions using both cofactors.

Model	Match genes	%	Mismatch genes	%	Both genes	%
iJO1366	65	79.27	9	10.98	8	9.76
iNJ661m	40	65.57	15	24.59	6	9.84
Yeast 7.6	60	77.92	14	18.18	3	3.90
iAbaylyiv4	57	78.08	16	21.92	0	0.00
iAI558	24	61.54	10	25.64	5	12.82

iBT721	37	68.52	17	31.48	0	0.00
iCG230	17	85.00	3	15.00	0	0.00
iCM925	33	61.11	11	20.37	10	18.52
iCR744	37	64.91	18	31.58	2	3.51
iCyc792	35	58.33	7	11.67	18	30.00
iCyj826	33	62.26	5	9.43	15	28.30
iCyn731	28	60.87	6	13.04	12	26.09
iCyp752	31	67.39	6	13.04	9	19.57
iJB785	26	74.29	8	22.86	1	2.86
iJL480	19	82.61	3	13.04	1	4.35
iJS747	33	68.75	14	29.17	1	2.08
iMF721	44	67.69	10	15.38	11	16.92
iMG746	19	61.29	2	6.45	10	32.26
iMP240	17	80.95	4	19.05	0	0.00
iNF518	32	80.00	8	20.00	0	0.00
iPS189	6	85.71	1	14.29	0	0.00
iRR1083	60	75.95	17	21.52	2	2.53
iWZ663	28	71.79	7	17.95	4	10.26
iTM560	28	73.68	5	13.16	5	13.16
iCac802	24	60.00	8	20.00	8	20.00
iCyt773	27	60.00	8	17.78	10	22.22
iEM439	17	70.83	4	16.67	3	12.50
iJH728	26	65.00	11	27.50	3	7.50
iJP962	61	76.25	11	13.75	8	10.00
iPB890	48	60.76	9	11.39	22	27.85
iSO783	46	85.19	8	14.81	0	0.00
iMK1208	73	80.22	12	13.19	6	6.59
iMO1056	67	84.81	10	12.66	2	2.53
iMP429	19	73.08	6	23.08	1	3.85
iMZ1055	64	63.37	23	22.77	14	13.86
iNV706	31	79.49	8	20.51	0	0.00
iRM588	30	65.22	14	30.43	2	4.35
iRsp1140	63	79.75	15	18.99	1	1.27
iJL846	35	66.04	14	26.42	4	7.55
iJSPpropionicus	3	75.00	1	25.00	0	0.00
iYLW1028	66	68.04	22	22.68	9	9.28
iJSPcarbinolicus	39	62.90	14	22.58	9	14.52

iKY620	27	71.05	8	21.05	3	7.89
iAM388	20	86.96	3	13.04	0	0.00
iMR539	19	73.08	5	19.23	2	7.69
iCyh755	30	69.77	4	9.30	9	20.93
iYS432	15	88.24	2	11.76	0	0.00

The above displayed results show the amount of predicted genes that match the cofactor used in their reactions. It is also displayed the amount of genes that are associated to reactions using NAD(H) but also to reactions that use NADP(H). In average, across all models, 71.9% of the predicted genes match the cofactor used in their encoded reactions. GEM *i*YS432 is the model with the highest percentage of matching genes, with 88.24%, while iCyc792 is the lowest, with 58.33% of the predicted genes mismatching cofactor usage.

As to predicted genes mismatching cofactor usage in all encoded reactions, *i*CR744 is the GEM with the highest percentage of mismatches with 31.6%, while *i*MG746 has the lowest, with 6.45%. When analyzing the distribution of genes in GPRs of reactions using NAD(H) and also reactions using NADP(H), we encounter *i*MG746 with the highest percentage, with 32.26%. From the analyzed 47 GEMs, 11 models do not possess any predicted gene encoding reactions using both cofactors separately.

Being *E. coli* and *S. cerevisiae* generally regarded as the best characterized microorganisms, an effort was made to further analyze the cases where the predicted gene cofactor specificity mismatched the gene's reaction cofactor usage in the models, or cases where reactions associated to a gene used both cofactors. To do that, experimental evidence was sought in the literature in an attempt to prove the correct cofactor specificity of the analyzed genes from the latest model updates of each organism, iJO1366 and Yeast 7.6.

A total of 65 *E. coli*'s *i*JO1366 predicted genes matched reaction cofactor usage, while 9 did not and 8 encoded for reactions using both cofactors. Yeast 7.6 had 60 predicted genes matching reaction cofactor usage, while 14 did not. Also, 3 of the predicted genes encoded reactions using NAD(H) and NADP(H).

Despite being two of the best characterized organisms, several references in the literature attribute a different cofactor specificity or state an unclear cofactor usage, in some of the genes predicted with a mismatching cofactor. In the case of *E. coli*, evidence supporting the cofactor specificity indicated by our tool was found for two of the mismatching genes and for all genes associated to reactions using both cofactors. Gene b1288, *FabI*, was characterized as NAD(H) dependent [92], despite being associated to 26 different reactions in *i*JO1366, 14 of which using NADP(H). Other cases, where dual specificity is claimed, were found to have a very strong preference for one of the cofactors; a feature that is not depicted in the model, and might indicate that in normal conditions, the enzyme would prefer the cofactor for which it has a higher affinity. Genes b1300 [93], b1525 [94], b0312 [95], b4267 [96] and b2552 [97] are associated in the model with reactions using both cofactor. However, literature reports a much higher affinity of their encoding enzymes for NAD(H) than for NADP(H). On the other hand, b1033 [98] and b3553 [99] are reported as highly preferring NADP(H) over NAD(H). Other cases of mismatching genes, associated to reactions using only the opposite cofactor are also unclear. For example b2040 [100] and b3608 [101], which encode in the model to NADP(H) specific reactions, were characterized as accepting both cofactors and predicted by our tool as NAD(H) specific. Gene b4266's enzyme is claimed to reduce 5-ketogluconate to D-gluconate using either NADH or NADPH. Nonetheless, the enzyme can only oxidize D-gluconate using NAD(H), with the use of NADP(H) resulting in lower specificity [96]. Despite the model assuming NADP(H) in this reaction, our prediction indicates NAD(H) specificity for this enzyme.

When analyzing the prediction mismatches on the genes corresponding to model Yeast 7.6, search in the literature supported the prediction for 6 of the genes with reactions using the opposite cofactor and the three genes associated to reactions using both cofactors. Gene YDR376W, associated to a reaction using NAD(H) in Yeast 7.6 has been found to have NADP-dependent reductase activity [102]. Concerning gene YER0773W, which is associated to two reactions using NADP(H), it has been claimed to use both cofactors- Nonetheless, Wang and coworkers found that, in the presence of potassium, NAD(H) reaction is favored in a much higher fold [103]. Gene YBR006W, associated to NADP(H) in the model, is claimed as having 2.5 fold higher activity using NAD(H) [104]. There are also some genes that, despite not proven in the literature, are characterized in Uniprot as NAD(H) specific, while associated to reactions using NADP(H) in the model. Gene YDR127W, YPL023C and YGL125W are all stated as binding to NAD(H) in this database. As to the genes associated to reactions using both cofactors, YJR139C was proven to use NAD(H) [105]. Gene YOR374W, which also catalyzes the reactions associated to YER073W, using NADP(H), is also associated to reactions using NAD(H) and considered NAD(H) dependent [103]. Gene YGL001C that is also associated to reactions using both cofactors has been experimentally determined as using NAD(H) [106]. In table 4.4 the number of reactions affected with cofactor misusage, by model are displayed.

Table 4.4 – GEM distribution of the total amount of reactions with cofactor usage matching the predicted genes, as well as reactions mismatching the predicted genes cofactor and reactions with GPR composed of predicted genes with both cofactors.

Model	Mismatching reactions	%	Matching reactions	%	Reactions with GPR with both cofactors
iJO1366	34	22.37	118	77.63	0
iNJ661m	50	50.00	50	50.00	8
Yeast 7.6	18	14.29	108	85.71	0
iAbaylyiv4	13	12.15	94	87.85	5
iAI558	24	42.86	32	57.14	7
iBT721	14	30.43	32	69.57	4
iCG230	9	32.14	19	67.86	6
iCM925	24	29.63	57	70.37	1
iCR744	27	32.14	57	67.86	4
iCyc792	54	39.71	82	60.29	22
iCyj826	44	34.92	82	65.08	17
iCyn731	39	31.45	85	68.55	0
iCyp752	19	21.59	69	78.41	0
iJB785	18	30.51	41	69.49	0
iJL480	4	12.50	28	87.50	0
iJS747	16	24.62	49	75.38	1
iMF721	69	48.94	72	51.06	28
iMG746	13	26.53	36	73.47	0
iMP240	4	14.29	24	85.71	0
iNF518	9	15.52	49	84.48	1
iPS189	1	11.11	8	88.89	0
iRR1083	26	26.80	71	73.20	10
iWZ663	19	32.20	40	67.80	7
iTM560	23	28.05	59	71.95	2
iCac802	14	17.28	67	82.72	3
iCyt773	18	25.35	53	74.65	1
iEM439	20	29.85	47	70.15	0
iJH728	14	23.73	45	76.27	1
iJP962	24	20.69	92	79.31	15
iPB890	48	27.59	126	72.41	24
iSO783	25	30.12	58	69.88	19
iMK1208	116	50.00	116	50.00	58
iMO1056	12	14.63	70	85.37	7

iMP429	8	22.86	27	77.14	2
iMZ1055	36	33.96	70	66.04	18
iNV706	16	37.21	27	62.79	12
iRM588	21	31.34	46	68.66	7
iRsp1140	24	23.08	80	76.92	12
iJL846	19	38.78	30	61.22	6
iJSPpropionicus	1	20.00	4	80.00	0
iYLW1028	73	49.66	74	50.34	38
iJSPcarbinolicus	16	22.54	55	77.46	1
iKY620	35	50.72	34	49.28	12
iAM388	7	21.21	26	78.79	5
iMR539	8	21.62	29	78.38	1
iCyh755	14	17.28	67	82.72	0
iYS432	2	11.11	16	88.89	0

In the displayed data the number of reactions are shown, in each model, using the incorrect cofactor, according to our tool. In average, 72.2% of the reactions analyzed, using NAD(P)(H), across all models are in agreement with the cofactor prediction of their encoding genes; nonetheless, these values vary quite substantially when analyzing each model individually.

GEM *i*MK1208 is the model most affected by reactions with wrongly attributed cofactor specificity, with 116 of the analyzed reactions having genes with a different cofactor specificity predicted. However, by percentage of analyzed genes, *i*KY620 is the GEM with the highest percentage of affected reactions, with 50.7% (35) of the 69 reactions analyzed having genes with a different cofactor specificity. GEM *i*PS189, in the other hand, is the model with the highest agreement between predicted cofactor and the reaction cofactor usage, with 88.9% (9) matching reactions.

GEMs *i*JO1366 and Yeast 7.6 respectively have 22.4% (34) and 14.3% (18) of their reactions using a different cofactor than the one predicted. The fact that these values are below the average 27.8% of affected reactions might possibly be due to being two of the best studied organisms, hence having their genome information well curated. Several models also have reactions that have in their GPR genes with different cofactor specificity, indicating erroneous GPR reconstructions, once genes that are specific for one cofactor are associated to reactions using the other cofactor. A possible solution for such cases is to split the reaction in two different reactions, one with NAD(H) as cofactor and the other with NAD(P)(H), and separate the genes in the GPR by cofactor specificity However, such task should be performed only when reliable information is available in order to prevent the potential formation of futile cycles.

4.3.4 Comparison of genome-scale metabolic models from *Saccharomyces cerevisiae*

S. cerevisiae is one of the best known and most widely studied microorganisms in science and was naturally among the first organisms having their metabolism mathematically represented through the reconstructions of its genome scale metabolic model [107]. This fact, associated with the ever developing amount of tools and techniques for genome analysis and metabolic characterization [5], [108], gave origin to multiple GEMs reconstructions [109]. Here, six *S. cerevisiae* GEMs were analyzed, having their origin and inspiration associated to two original GEM reconstructions, GEM *i*FF708, published in 2003 [107], and the Yeast consensus model, first published in 2008 [110]. GEM *i*IN800, the oldest GEM analyzed in this study, published in 2008 [37], was developed as an updated version of the three previously existing models iFF708, iND750 [111] and iLL672 [112]. GEM *i*MM904 [39], developed in 2009, also updates iND750, while *i*AZ900 [36], published in 2010 is itself an updated version of *i*MM904. GEM *i*TO977 [41], published in 2013, is based on *i*IN800 and the consensus model, and then improved and expanded using gap-filling methods and by introducing new reactions and pathways based on studies of the literature and databases. Yeast 6 [42] was published in 2013 and is a continuous revision of the original yeast consensus model, while Yeast 7.6 [33], the most recent and best curated model, started as Yeast 7.00, with a revision of the representation of fatty acid, glycerolipid, and glycerophospholipid metabolism having Yeast 6 as the basis, but has been frequently updated since, with the most recent version (Yeast 7.6) being released in 2015. Among all models, there is a total of169 different genes related to reactions using NAD(P)(H). Figure 4.4 displays the gene distribution among the analyzed models, with information on model's total gene and reaction count, as well as total amount of genes and reactions associated to NAD(P)(H).

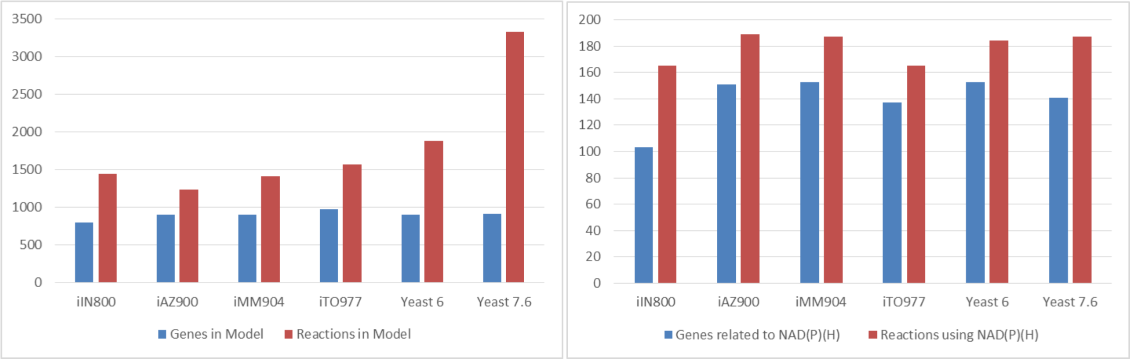


Figure 4.4 – In the left is depicted a comparison between the total amount of genes and reactions included in each GEM. On the right, a comparison between the total amount of genes and reactions using NAD(P)(H) included in each GEM.

In the displayed figure it is possible to see the distribution of the number of genes and reactions among the six GEMs analyzed. GEM *i*IN800, the earliest developed GEM, has the lowest amount of genes and also the least amount of genes associated to NAD(P)(H). The amount of total genes present remains relatively similar across the developed models with an average of 899 genes, while the amount of total reactions ranges from 1240 reactions, for *i*AZ900, to 3334 reactions for Yeast 7.6. The justification for this discrepancy lies in the inclusion of membrane compartments in the model, to enhance more resolution on reaction localization, thus requiring the addition of transport reactions, heavily increasing the number of added reactions in the model [33]. As to the number of genes and reactions related to NAD(P)(H), model updates do not appear to influence gene or reaction number, with the latest models developed having a smaller gene count when compared with the others, with the exception of *i*IN800. GEM *i*IN800 and *i*TO977 have the lowest amount of reactions related to NAD(P)(H), with 165 reactions each, while the remaining models have a reactions count ranging from 184 to 189.

In order to acquire a better insight on the evolution of model curation of reactions and genes related to the usage of NAD(P)(H), along the years, a comparison between the presented models was performed. Firstly, GEMs iIN800, iAZ900, iMM904 and iTO977 were compared between them and the differences analyzed, while Yeast 6 and Yeast 7.6 were compared against each other. In the end, iTO977 and Yeast 7.6, the last developed models were also compared.

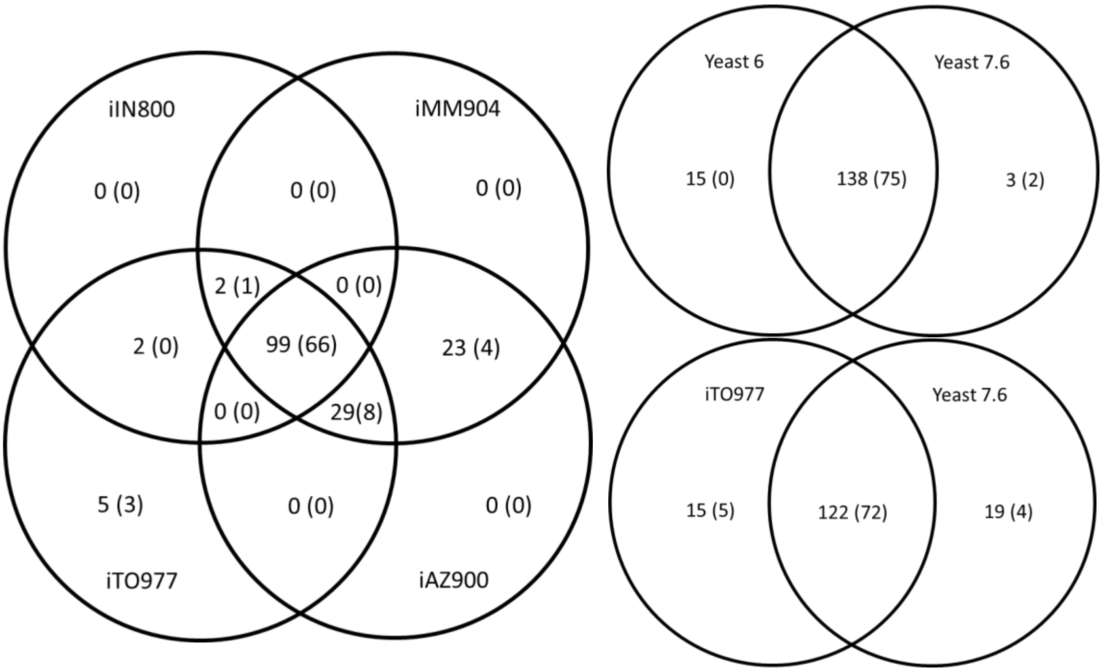


Figure 4.5 – Veen diagrams with the intersection of the common genes related to NAD(P)(H) contained in each model and also, in brackets, the intersection of the common predicted genes related to NAD(P)(H) contained in each model.

The diagrams presented in Figure 4.5 display the number of genes related to NAD(P)(H) that each model contains. In brackets is displayed the amount of predicted genes. When analyzing the displayed data in the diagram correspondent to the four models we can see that these models share 99 genes related to NAD(P)(H), from which 66 had their cofactor specificity predicted. GEM *i*TO977 is the only model with genes that are not shared with any of the previously developed models, with 5 genes. GEM *i*MM904 and *i*AZ900, that were developed as updates of *i*MM904, share 23 genes that are not present in the other two models, having only 4 of these been predicted for cofactor specificity. When analyzing the yeast consensus models, Yeast 6 and Yeast 7.6, we can see that these models share 138 genes, being 75 predicted. Despite Yeast 6 having 15 genes that are not present in Yeast 7.6, none of them have their cofactor predicted, while two of the three genes in Yeast 7.6 that are not present in Yeast 6 have a prediction made. When comparing the genes shared by the two latest models, iTO977 and Yeast 7.6, we can see that these models share 122 genes, with 72 having a prediction performed. GEM *i*TO977 incorporates 15 genes that are not present in Yeast 7.6, being 5 predicted, while Yeast 7.6 possesses 19 genes that are not present in *i*TO977, being 4 predicted.

The results achieved by the developed software are displayed in Figure 4.6. Gene cofactor prediction affects directly the information on the gene, but also the information on the reactions that have those genes present in the respective GPR. The displayed figures show the amount of predicted genes, for each model, that match or mismatch cofactor utilization. Situations where a specific gene is associated to GPRs from reactions using both cofactors are also highlighted. Also shown is the amount of reactions, per model, whose cofactor utilization matches, or mismatches, the cofactor predictions of the genes represented in its GPR.

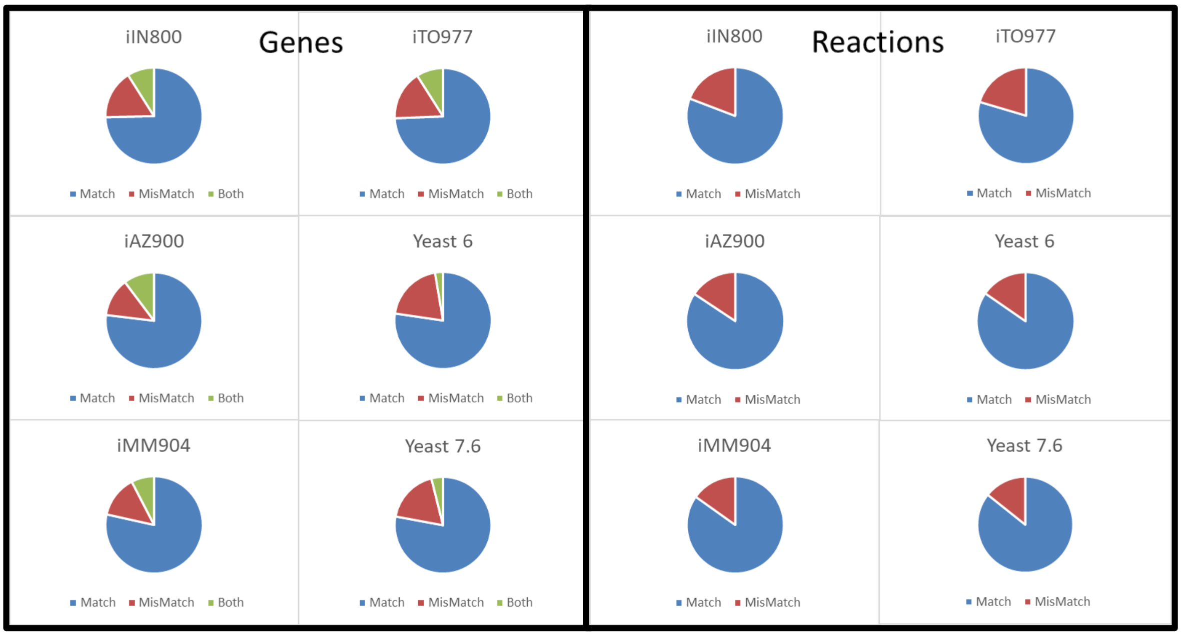


Figure 4.6 – Comparison between the six analyzed *S. cerevisiae* GEM. Left: Gene prediction comparison, with displayed data showing the amount of matching and mismatching predicted genes with reaction cofactor usage as well as genes present in reactions using both cofactors. Right: Reaction matches comparison displaying the amount of reactions, in each model, that match or mismatch the cofactor predicted.

When analyzing the displayed figures we can see that the vast majority of the predicted genes match the cofactor usage by the reactions. Between 74.4% and 78.5% of the predicted genes match cofactor utilization of their related reactions, being *i*MM904 the GEM with the highest percentage of matching genes, with 62, while *i*IN800 has the lowest percentage of matching genes, with 50%. As to the predicted genes mismatching reaction cofactor, Yeast 6 has the highest percentage with 20% (15 genes), while *i*AZ900 has the lowest, with 10 genes. There are also genes that are present in GPR from reactions using NAD(H) and NAD(P)(H). GEM *i*AZ900 is the model with the highest percentage of genes, with 10.3% (8 genes), which helps justifying the lower amount of mismatching genes. The lowest percentage of these genes are present in Yeast 6, with 2.7% (2 genes) of genes being present in GPR of reactions using both cofactors.

An interesting show case on the applicability of the developed methodology is the cofactor prediction of the gene YHR037W. Gene YHR037W, which encodes the mitochondrial enzyme Delta-1-pyrroline-5-carboxylate dehydrogenase and is responsible for three different reactions, L-1-pyrroline-3-hydroxy-5-carboxylate dehydrogenase, L-4-hydroxyglutamate semialdehyde dehydrogenase and 1-pyrroline-5-carboxylate dehydrogenase, has been characterized as possessing specificity for NAD(H) usage as cofactor [113], [114]. This enzyme was also predicted as being specific for NAD(H) by our software, which is in accordance with the data found in the literature. Nonetheless, this fact is only true in the models from the consensus yeast model, Yeast 6.0 and Yeast 7.6. All other models analyzed use reactions associated to this enzyme with NADP(H) as cofactors. The fact that the consensus yeast models represent the most recent update versions of *S. cerevisiae* GEMs might explain the models better curation, in this case. This analysis shows once again the advantages and importance of the developed software in genome-scale metabolic model reconstruction.

4.3.5 Analysis of *S. cerevisiae* Yeast 7.6 GEM

In order to assess the applicability of the results achieved by NiCofactor, the latest *S. cerevisiae* model Yeast 7.6 was further investigated. Mismatching genes were individually inspected and also the corresponding reactions were corrected, according to the developed tool, and the model was used in simulations in order to attest the implication of the new set of curated reactions in the simulation results. As previously referred, Yeast 7.6 is the latest model developed for *S. cerevisiae*, and represents the metabolism of *S. cerevisiae* s288c.

As previously mentioned for Yeast 7.6, approximately 18.2% of the cofactor predicted genes do not match the cofactor used by the reactions encoded by them, while 3.9% of the analyzed genes are present in GPRs of reactions using both cofactors separately. In table 4.5 the genes that do not match the reactions cofactor usage are displayed, as well as the corresponding reactions.

Table 4.5 – Predicted genes with mismatching cofactor usage and the respective reactions.

Gene name	Mismatching reactions	Reaction Cofactor	Cofactor prediction
YDL215C	r_0470	NAD(H)	NADP(H)
YDR376W	r_0530	NAD(H)	NADP(H)
YJR139C	r_0547	NADP(H)	NAD(H)
YGL001C	r_0234	NADP(H)	NAD(H)
YOR374W	r_0175; r_0178	NADP(H))	NAD(H)
YMR041C	r_0320	NAD(H)	NADP(H)

YER073W	r_0175; r_0178	NADP(H)	NAD(H)
YBR115C	r_0678	NADP(H)	NAD(H)
YPL061W	r_0177; r_0173	NADP(H)	NAD(H)
YJR137C	r_1027	NADP(H)	NAD(H)
YBR006W	r_1023	NADP(H)	NAD(H)
YDR127W	r_0996	NADP(H)	NAD(H)
YGR204W	r_0732	NADP(H)	NAD(H)
YBR084W	r_0733	NADP(H)	NAD(H)
YPL023C	r_0080	NADP(H)	NAD(H)
YGL125W	r_0080	NADP(H)	NAD(H)
YLR355C	r_0669; r_0096	NADP(H)	NAD(H)

As it is possible to observe, a total of 17 genes are represented, which are included in GPRs of 18 reactions. Three of the genes in the table encode for reactions using NAD(H) in the model but were predicted as NADP(H) specific, while 14 encode for reactions using NADP(H) but were predicted as NAD(H) specific.

With the final set of genes predicted as specific for a different cofactor, the affected reactions present in model Yeast 7.6 were altered in their cofactor usage, giving origin to a new model: Yeast 7.6_Corrected. Table 4.6 displays the reactions affected by the corrections implemented to the model, and the type of corrections implemented in order to match the predicted cofactor specificity of the encoding genes. The majority of the corrections implemented consisted in the alteration of the cofactor used from NADP(H) to NAD(H). In the cases of existing reactions already using the opposite cofactor in the model, these were indicated for deletion to prevent the creation of duplicated reactions.

Table 4.6 – Yeast 7.6_ Corrected model with the description of the performed corrections along with reaction identification and reaction name.

Reaction ID	Reaction name	Correction
r_0320	D-arabinose 1-dehydrogenase (NAD)	Delete reaction
r_0530	Heme O monooxygenase	NAD(H) -> NADP(H)
r_0470	Glutamate dehydrogenase (NAD)	NAD(H) -> NADP(H)
r_0669	Ketol-acid reductoisomerase (2-aceto-2-hydroxybutanoate)	NADP(H) -> NAD(H)
r_0080	5,10-methylenetetrahydrofolate reductase (NADPH)	NADP(H) -> NAD(H)
r_0177	Aldehyde dehydrogenase (indole-3-acetaldehyde, NADP)	NADP(H) -> NAD(H)
r_0175	Aldehyde dehydrogenase (acetylaldehyde, NADP)	Delete reaction

r_0178	Aldehyde dehydrogenase (indole-3-acetaldehyde, NADP)	Delete reaction
r_0547	Homoserine dehydrogenase (NADP)	Delete reaction
r_0173	Aldehyde dehydrogenase (acetaldehyde, NADP)	NADP(H) -> NAD(H)
r_0234	C-3 sterol dehydrogenase	NADP(H) -> NAD(H)
r_0096	Acetohydroxy acid isomeroreductase	NADP(H) -> NAD(H)
r_1027	Sulfite reductase (NADPH2)	NADP(H) -> NAD(H)
r_1023	Succinate-semialdehyde dehydrogenase (NADP)	NADP(H) -> NAD(H)
r_0678	L-aminoadipate-semialdehyde dehydrogenase (NADPH)	NADP(H) -> NAD(H)
r_0733	Methylenetetrahydrofolate dehydrogenase (NADP)	NADP(H) -> NAD(H)
r_0996	Shikimate dehydrogenase	NADP(H) -> NAD(H)
r_0732	Methylenetetrahydrofolate dehydrogenase (NADP)	Delete reaction

In order to assess the effect of these changes in cofactor usage on the overall performance of the model, a comparison between the original and corrected models was employed. Both models were set and simulated with the same environmental conditions, being their biomass growth representation, and flux distribution in key metabolic routes of the central metabolism, compared and evaluated.

The starting point of any metabolic simulation analysis is the observation of the maximum biomass growth rate achieved under predetermined environmental conditions. Being the experimental *S. cerevisiae* biomass yield on glucose around 0.5gCDW/g Glucose [115] and the maximum glucose uptake rate on the model's environmental conditions set to 1.15 mmol/(gCDW.h), the expected growth rate should be close to 0.10 h^{-1}. Both model simulations predicted a maximum biomass growth rate similar to the expected value, with the original Yeast 7.6 achieving 0.1089 h^{-1}, and the corrected version 0.1075 h^{-1}. However, when analyzing the metabolic fluxes in the central metabolism, some inconsistencies were detected, most noticeably, in the Pentose Phosphate Pathway (PPP), where there was an almost absent flux, in both models. Flux in the oxidative phase of this pathway is required for the generation of NADPH in the cell, and its deficiency indicates that the model is artificially generating NADPH somewhere else. A similar problem was found in a recent study involving older *S.cerevisiae* models. In the performed analysis, Pereira and coworkers [116] simulated the GEMs *i*FF708, *i*MM904, *i*TO977 and Yeast 6 and found that the analyzed models were predicting erroneous fluxes in central carbon pathways, especially in the pentose phosphate pathway. Upon further investigation it was found that among the problems with the simulations were the consumption and production of NAD(P)(H). After intense literature review and manual curation of these reactions, a set of reactions was forcibly constrained or inactivated, with more accurate flux distributions being simulated by the models. In the analyzed Yeast 7.6 model, upon further investigation it was detected that, in the original model, NADPH was being produced by the cytosolic Methylenetetrahydrofolate dehydrogenase (r_0732). Besides artificially generating NADPH, the high activity of this enzyme also originated a high flux in the Folate pathway, creating a circular flux of consumption and production of Tetrahydrofolate, while producing ATP in reaction r_0446. This inconsistency is solved in the corrected model with the cofactor change of r_0732 from NAD(P)(H) to NAD(H), dramatically decreasing its flux. Despite the undergone modifications, in the corrected model there were still some inconsistencies, as the flux was not yet restored in the PPP. The responsible for this was the cytosolic isocitrate dehydrogenase (r_0659), encoded by the gene YLR174W, first characterized by [117], which was now supplying the cell with NADPH, although with lesser flux. This gene, however, has been experimentally characterized as being repressed in the presence of glucose [118]. Also, the activity of this enzyme has been experimentally shown to be inhibited when the concentration of NADPH increases relatively to NADP$^+$ [119]. This increase in concentration might also play a role in the redirection of Gibbs free energy in this reaction, facilitating NADP$^+$ production *in vivo*. For all these reasons, a decision was taken to constrain the reversibility of r_0659 in the direction of NADPH consumption. With this small change in both models, we were able to observe a dramatic flux change in the corrected model with the total restoration of the PPP. Central carbon metabolism fluxes from both models, along with experimental flux determinations retrieved from the literature [120], [121] are displayed in Figure 4.7.

A detailed analysis of the presented figure reveals the astonishing improvement in metabolic flux on the Pentose Phosphate Pathway, as well as in the citric acid cycle, in the corrected Yeast 7.6 model, when compared to the original model. Also, when comparing model flux and experimental data, Yeast 7.6_corrected reveals itself to be a very close approximation to the *in vivo* metabolic fluxes occurring inside the cell.

The oxidative phase of PPP (oxPPP), represented in the figure by the Phosphogluconate dehydrogenase (reaction GND, r_0889), shows a rather slender flux in the original model, as it is regenerating NADPH using the Methylenetetrahydrofolate dehydrogenase (reaction r_0732).

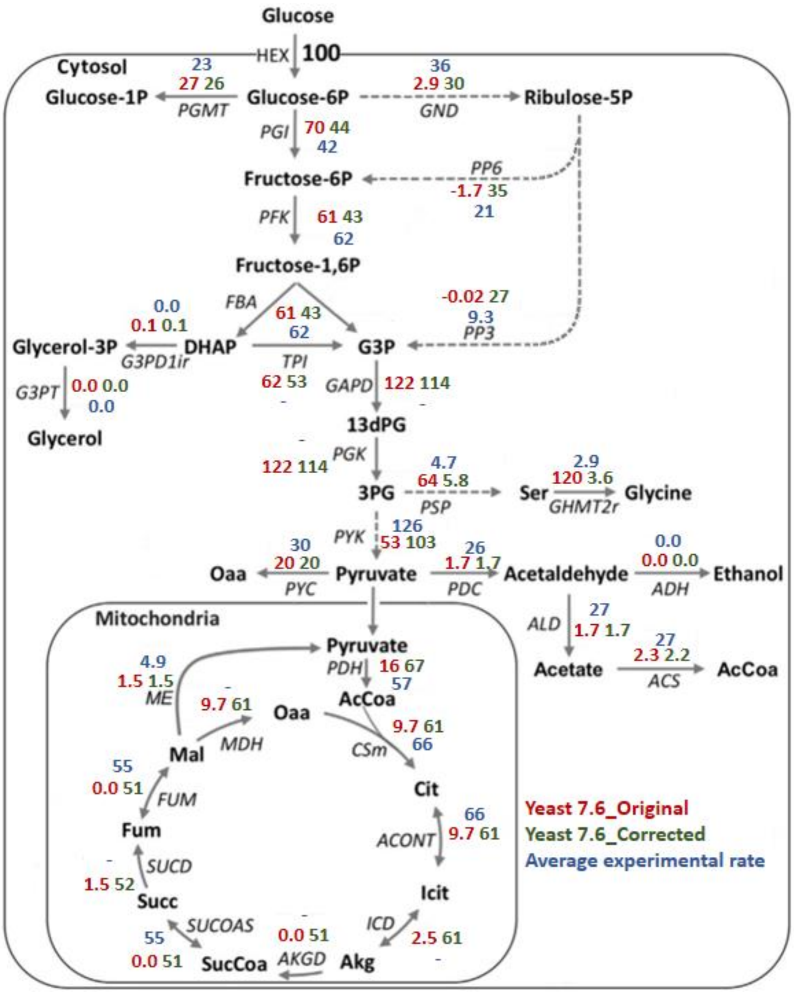


Figure 4.7 - Comparison between the predictions of pFBA for the original and corrected models of Yeast 7.6 and also with experimental data. Original Yeast 7.6 (red), Corrected Yeast 7.6 (Green), Average of the 13C-MFA fully aerobic chemostat at a dilution rate of 0.1 h^{-1} (blue). Reactions: ACONT- aconitase, ACS- acetyl-CoA synthetase, ADH- alcohol dehydrogenase, AKGD- alpha-ketoglutarate dehydrogenase, ALD- aldehyde dehydrogenase, CSm- citrate synthase, FBA- fructose 1,6-bisphosphate aldolase, FUM-fumarase, G3PD1ir- glycerol-3-phosphate dehydrogenase, G3PT- glycerol-1-phosphatase, GAPD- glyceraldehyde-3-phosphate dehydrogenase, GHMT2r- serine hydroxymethyltransferase, GND- 6-phosphogluconate dehydrogenase, HEX- hexokinase, ICD-mitochondrial isocitrate dehydrogenase, MDH- mitochondrial malate dehydrogenase, MEmitochondrial malic enzyme, PDC- pyruvate decarboxylase, PDH- pyruvate dehydrogenase, PFKphosphofructokinase, PGI- phosphoglucose isomerase, PGK- 3-phosphoglycerate kinase, PGMTphosphoglucomutase, PP3- sum of the non-oxidative reactions of the pentose phosphate pathway producing glyceraldehyde-3-phosphate, PP6- sum of the non-oxidative reactions of the pentose phosphate pathway producing fructose-6-phosphate, PSP- phosphoserine phosphatase, PYC- pyruvate carboxylase, PYKpyruvate kinase, SUCD- succinate dehydrogenase, SUCOAS- succinyl-CoA ligase, THRS- threonine synthase, TPI- triose phosphate isomerase. Metabolites: 13dPG- 1,3-diphosphoglycerate, 3PG- 3-phosphoglycerate, AcCoA- acetyl-CoA, Akg- 2-oxoglutarate, Cit- citrate, DHAP- dihydroxyacetonephosphate, Fum- fumarate, G3P- glyceraldehyde-3-phosphate, Icit- isocitrate, Mal - L-malate, Oaaoxaloacetate, Ser- L-serine, Succ- succinate, SucCoa- succinyl-CoA.

In the corrected model, the oxPPP is completely restored, encompassing flux values very similar to the fluxes retrieved from experimental data. The non-oxidative phase is represented in the figure by the aggregation of reactions leading to the production of Fructose-6-phosphate, in reaction PP6, and Glyceraldehyde-3-phosphate, in reaction PP3. In these reactions, once again, the corrected model shows satisfactory results, although this time with fluxes smaller than in literature. Meanwhile, the original model, not only is unable to complete the pathway, but actually reverts its fluxes. This happens due to the absence of flux in the oxPPP, as the intermediates in need for biomass production are produced from the products of glycolysis. Moving along in metabolism, fluxes from Phosphoserine phosphatase (reaction PSP, r_0917) producing L-serine and Glycine hydroxymethyltransferase producing L-glycine (reaction GHMT2r, r_0502), are presented with very high values in the original model. The reason behind these elevated fluxes is the artificial regeneration of NADPH occurring in the folate pathway, which requires the consumption of these two aminoacids. In the corrected model, these fluxes are according to the literature, which favor a higher flux in the Pyruvate kinase (reaction PYK, r_0962), as suggested in the experimental data. The higher flux of PYK, when compared to the original model, translates in a higher availability of pyruvate to enter mitochondria and help complete the TCA cycle. Fluxes from the TCA cycle in both models are also quite disparate, as the corrected model fluxes in this pathway show a complete cycle and are again very similar to the average values found in the literature. On the other hand, TCA cycle fluxes in the original model are overall smaller, and even absent in some reactions.

When analyzing the fluxes in the altered cofactor reactions, in the corrected model, it is possible to observe that, overall, the implemented modifications had an effect in the increase of NAD$^+$ availability, while decreasing NADP$^+$ production. From the set of 18 reactions that had their cofactor modified to match the encoding genes cofactor specificity, 10 were found to have fluxes when simulated with the specified environmental conditions. From these, only r_0530 (heme O monooxygenase) had a modification that produce NADP$^+$, with a flux close to zero. All other reactions with flux started producing NAD$^+$ instead of NADP$^+$, with the exception of r_0173 (Aldehyde dehydrogenase) and r_0996 (Shikimate dehydrogenase) that produced NADH, instead of NADPH. One other detail, in this set of reactions, is that all reactions with fluxes maintained a similar flux when the corresponding cofactor was modified. An example of this is the aldehyde dehydrogenase represented in figure 4.7 by ALD that maintained the same flux, despite using a different cofactor. The only exception to this is again r_0732 (Methylenetetrahydrofolate dehydrogenase) that, due to the inversion of fluxes on PPP, saw its flux (in the equivalent reaction r_0731, that already used NAD(H)) decreasing almost 40 fold.

The analyzed results, from the fluxes in the central carbon metabolism, suggest a better depiction of reality from the corrected model, when directly compared with the original one. The changes in cofactor usage in the selected reactions have proven to provide the model with a far more accurate result, when predicting metabolic fluxes. Experimental evidence on central carbon metabolic flux distribution and cofactor specificity also support the implemented corrections to the original model Yeast 7.6. The implementation of the used tool for automatically performing cofactor specificity prediction revealed of upmost utility in the achievement of the discussed results.

4.4 Conclusions

Genome-scale metabolic model reconstructions are a laborious and highly skill requiring tasks that demands a great effort in the curation process. Reaction cofactor prediction in GEM reconstruction has an immediate effect on reaction composition. With the usage of NAD(P)(H) being scrutinized, the association of a cofactor to a specific gene determines that the reactions associated to that gene should be using the gene associated cofactor. In the presented study, the cofactor specificity of several enzymes was assessed in order to analyze and state of NAD(P)(H) cofactor usage curation within several GEM. The aminoacid sequence of each enzyme associated to reactions using NAD(P)(H) was retrieved and processed using our developed tool, NiCofactor. Results evidence an overall satisfactory curation of NAD(P)(H) usage in the analyzed models, despite the occurrence of some mismatches that might impair an accurate GEM simulation. An emphasis was given to GEMs modeling *S. cerevisiae* metabolism due to their abundance and relevance. The most recent yeast GEM, Yeast 7.6 was further analyzed and its reactions corrected, according to our data. Both original and corrected models were simulated in identical conditions, with surprising results achieved by the corrected model when compared not only with the original one, but also with the literature. This work demonstrates the performance and applicability of the developed software in the curation of GEMs, exhibiting a great potential as a tool for aiding GEM reconstruction for a large group of researches.

4.5 References

[1] J. Nielsen, "Metabolic engineering: Techniques for analysis of targets for genetic manipulations," *Biotechnol. Bioeng.*, vol. 58, no. 2–3, pp. 125–132, 1998.

[2] R. Ledesma-Amaro, E. J. Kerkhoven, J. L. Revuelta, and J. Nielsen, "Genome scale metabolic modeling of the riboflavin overproducer Ashbya gossypii," *Biotechnol. Bioeng.*, vol. 9999, no. xxx, pp. 1191–1199, 2014.

[3] J. E. Bailey, "Toward a Science of Metabolic Engineering," *Science (80-.).*, vol. 252, pp. 1668–1697, 1991.

[4] J. E. BAILEY, S. BIRNBAUM, J. L. GALAZZO, C. KHOSLA, and J. V. SHANKS, “Strategies and Challenges in Metabolic Engineering,” *Ann. N. Y. Acad. Sci.*, vol. 589, no. 1, pp. 1–15, 1990.

[5] B. M. Woolston, S. Edgar, and G. Stephanopoulos, “Metabolic Engineering: Past and Future,” *Annu. Rev. Chem. Biomol. Eng.*, vol. 4, no. 1, pp. 259–288, 2013.

[6] W. Tian and J. Skolnick, “How well is enzyme function conserved as a function of pairwise sequence identity?,” *J. Mol. Biol.*, vol. 333, no. 4, pp. 863–882, 2003.

[7] D. Eisenberg, E. M. Marcotte, I. Xenarios, and T. O. Yeates, “Protein function in the post-genomic era.,” *Nature*, vol. 405, no. 6788, pp. 823–6, 2000.

[8] H. Kitano, “Systems Biology: A Brief Overview,” *Science (80-.).*, vol. 295, no. 5560, pp. 1662–1664, 2002.

[9] A. M. Feist, M. J. Herrgård, I. Thiele, J. L. Reed, and B. Palsson, “Reconstruction of biochemical networks in microorganisms,” *Nat. Rev. Microbiol.*, vol. 7, no. 2, pp. 129–143, 2009.

[10] J. W. Lee, D. Na, J. M. Park, J. Lee, S. Choi, and S. Y. Lee, “Systems metabolic engineering of microorganisms for natural and non-natural chemicals,” *Nat. Chem. Biol.*, vol. 8, no. 6, pp. 536–546, 2012.

[11] J. M. Otero, D. Cimini, K. R. Patil, S. G. Poulsen, L. Olsson, and J. Nielsen, “Industrial Systems Biology of Saccharomyces cerevisiae Enables Novel Succinic Acid Cell Factory,” *PLoS One*, vol. 8, no. 1, pp. 1–10, 2013.

[12] C. Bro, B. Regenberg, J. Förster, and J. Nielsen, “In silico aided metabolic engineering of Saccharomyces cerevisiae for improved bioethanol production,” *Metab. Eng.*, vol. 8, no. 2, pp. 102–111, 2006.

[13] Z. L. Fowler, W. W. Gikandi, and M. A. G. Koffas, “Increased malonyl coenzyme A biosynthesis by tuning the Escherichia coli metabolic network and its application to flavanone production,” *Appl. Environ. Microbiol.*, vol. 75, no. 18, pp. 5831–5839, 2009.

[14] A. R. Brochado, C. Matos, B. L. Møller, J. Hansen, U. H. Mortensen, and K. R. Patil, “Improved vanillin production in baker’s yeast through in silico design,” *Microb. Cell Fact.*, vol. 9, pp. 1–15, 2010.

[15] M. A. Asadollahi, J. Maury, K. R. Patil, M. Schalk, A. Clark, and J. Nielsen, "Enhancing sesquiterpene production in Saccharomyces cerevisiae through in silico driven metabolic engineering," *Metab. Eng.*, vol. 11, no. 6, pp. 328–334, 2009.

[16] H. S. Choi, S. Y. Lee, T. Y. Kim, and H. M. Woo, "In silico identification of gene amplification targets for improvement of lycopene production," *Appl. Environ. Microbiol.*, vol. 76, no. 10, pp. 3097–3105, 2010.

[17] I. Rocha, J. Förster, and J. Nielsen, "Design and application of genome-scale reconstructed metabolic models.," *Methods Mol. Biol.*, vol. 416, pp. 409–431, 2008.

[18] I. Thiele and B. Palsson, "A protocol for generating a high-quality genome-scale metabolic reconstruction," *Nat Protoc*, vol. 5, no. 1, pp. 93–121, 2010.

[19] R. Agren, L. Liu, S. Shoaie, W. Vongsangnak, I. Nookaew, and J. Nielsen, "The RAVEN Toolbox and Its Use for Generating a Genome-scale Metabolic Model for Penicillium chrysogenum," *PLoS Comput. Biol.*, vol. 9, no. 3, 2013.

[20] J. J. Hamilton and J. L. Reed, "Software platforms to facilitate reconstructing genome-scale metabolic networks," *Environ. Microbiol.*, vol. 16, no. 1, pp. 49–59, 2014.

[21] R. A. Notebaart, F. H. J. van Enckevort, C. Francke, R. J. Siezen, and B. Teusink, "Accelerating the reconstruction of genome-scale metabolic networks," *BMC Bioinformatics*, vol. 7, pp. 1–10, 2006.

[22] C. S. Henry, M. DeJongh, A. a Best, P. M. Frybarger, B. Linsay, and R. L. Stevens, "High-throughput generation, optimization and analysis of genome-scale metabolic models.," *Nat. Biotechnol.*, vol. 28, no. 9, pp. 977–982, 2010.

[23] K. Arakawa, Y. Yamada, K. Shinoda, Y. Nakayama, and M. Tomita, "GEM system: Automatic prototyping of cell-wide metabolic pathway models from genomes," *BMC Bioinformatics*, vol. 7, pp. 1–11, 2006.

[24] R. U. Ibarra, J. S. Edwards, and B. O. Palsson, "*Escherichia coli* K-12 undergoes adaptive evolution to achieve *in silico* predicted optimal growth," *Nature*, vol. 420, no. 6912, pp. 186–189, 2002.

[25] N. E. Lewis, K. K. Hixson, T. M. Conrad, J. A. Lerman, P. Charusanti, A. D. Polpitiya, J. N.

Adkins, G. Schramm, S. O. Purvine, D. Lopez-Ferrer, K. K. Weitz, R. Eils, R. König, R. D. Smith, and B. Palsson, "Omic data from evolved E. coli are consistent with computed optimal growth from genome-scale models," *Mol. Syst. Biol.*, vol. 6, no. 390, 2010.

[26] D. Lee, O. Redfern, and C. Orengo, "Predicting protein function from sequence and structure," *Nat. Rev. Mol. Cell Biol.*, vol. 8, no. 12, pp. 995–1005, 2007.

[27] K. M. Borgwardt, C. S. Ong, S. Schonauer, S. V. N. Vishwanathan, A. J. Smola, and H.-P. Kriegel, "Protein function prediction via graph kernels," *Bioinformatics*, vol. 21, no. Suppl 1, pp. i47–i56, 2005.

[28] I. V. Tetko, I. V. Rodchenkov, M. C. Walter, T. Rattei, and H. W. Mewes, "Beyond the 'best' match: Machine learning annotation of protein sequences by integration of different sources of information," *Bioinformatics*, vol. 24, no. 5, pp. 621–628, 2008.

[29] P. D. Dobson and A. J. Doig, "Predicting enzyme class from protein structure without alignments," *J. Mol. Biol.*, vol. 345, no. 1, pp. 187–199, 2005.

[30] Z. A. King and A. M. Feist, "Optimizing Cofactor Specificity of Oxidoreductase Enzymes for the Generation of Microbial Production Strains—OptSwap," *Ind. Biotechnol.*, vol. 9, no. 4, pp. 236–246, 2013.

[31] C. N. Jensen, S. T. Ali, M. J. Allen, and G. Grogan, "Mutations of an NAD(P)H-dependent flavoprotein monooxygenase that influence cofactor promiscuity and enantioselectivity," *FEBS Open Bio*, vol. 3, pp. 473–478, 2013.

[32] C. N. Jensen, S. T. Ali, M. J. Allen, and G. Grogan, "Exploring nicotinamide cofactor promiscuity in NAD(P)H-dependent flavin containing monooxygenases (FMOs) using natural variation within the phosphate binding loop. Structure and activity of FMOs from Cellvibrio sp. BR and Pseudomonas stutzeri NF13," *J. Mol. Catal. B Enzym.*, vol. 109, pp. 191–198, 2014.

[33] H. W. Aung, S. A. Henry, and L. P. Walker, "Revising the Representation of Fatty Acid, Glycerolipid, and Glycerophospholipid Metabolism in the Consensus Model of Yeast Metabolism," *Ind. Biotechnol.*, vol. 9, no. 4, pp. 215–228, 2013.

[34] I. Rocha, P. Maia, P. Evangelista, P. Vilaça, S. Soares, J. P. Pinto, J. Nielsen, K. R. Patil, E.

C. Ferreira, and M. Rocha, “OptFlux: an open-source software platform for in silico metabolic engineering.,” *BMC Syst. Biol.*, vol. 4, p. 45, 2010.

[35] M. Hucka, A. Finney, H. M. Sauro, H. Bolouri, J. C. Doyle, H. Kitano, A. P. Arkin, B. J. Bornstein, D. Bray, A. Cornish-Bowden, A. A. Cuellar, S. Dronov, E. D. Gilles, M. Ginkel, V. Gor, I. I. Goryanin, W. J. Hedley, T. C. Hodgman, J. H. Hofmeyr, P. J. Hunter, N. S. Juty, J. L. Kasberger, A. Kremling, U. Kummer, N. Le Novère, L. M. Loew, D. Lucio, P. Mendes, E. Minch, E. D. Mjolsness, Y. Nakayama, M. R. Nelson, P. F. Nielsen, T. Sakurada, J. C. Schaff, B. E. Shapiro, T. S. Shimizu, H. D. Spence, J. Stelling, K. Takahashi, M. Tomita, J. Wagner, and J. Wang, “The systems biology markup language (SBML): A medium for representation and exchange of biochemical network models,” *Bioinformatics*, vol. 19, no. 4, pp. 524–531, 2003.

[36] A. R. Zomorrodi and C. D. Maranas, “Improving the iMM904 S. cerevisiae metabolic model using essentiality and synthetic lethality data,” *BMC Syst. Biol.*, vol. 4, no. 1, p. 178, 2010.

[37] I. Nookaew, M. C. Jewett, A. Meechai, C. Thammarongtham, K. Laoteng, S. Cheevadhanarak, J. Nielsen, and S. Bhumiratana, “The genome-scale metabolic model iIN800 of Saccharomyces cerevisiae and its validation: A scaffold to query lipid metabolism,” *BMC Syst. Biol.*, vol. 2, 2008.

[38] J. D. Orth, T. M. Conrad, J. Na, J. A. Lerman, H. Nam, A. M. Feist, and B. Palsson, “A comprehensive genome-scale reconstruction of Escherichia coli metabolism-2011,” *Mol. Syst. Biol.*, vol. 7, no. 535, pp. 1–9, 2011.

[39] M. L. Mo, B. Palsson, and M. J. Herrgård, “Connecting extracellular metabolomic measurements to intracellular flux states in yeast,” *BMC Syst. Biol.*, vol. 3, pp. 1–17, 2009.

[40] X. Fang, A. Wallqvist, and J. Reifman, “Development and analysis of an in vivo-compatible metabolic network of Mycobacterium tuberculosis,” *BMC Syst. Biol.*, vol. 4, no. 1, p. 160, 2010.

[41] T. Österlund, I. Nookaew, S. Bordel, and J. Nielsen, “Mapping condition-dependent regulation of metabolism in yeast through genome-scale modeling,” *BMC Syst. Biol.*, vol. 7, 2013.

[42] B. D. Heavner, K. Smallbone, N. D. Price, and L. P. Walker, “Version 6 of the consensus yeast metabolic network refines biochemical coverage and improves model performance,”

Database, vol. 2013, no. January, pp. 1–5, 2013.

[43] M. Durot, F. Le Fèvre, V. de Berardinis, A. Kreimeyer, D. Vallenet, C. Combe, S. Smidtas, M. Salanoubat, J. Weissenbach, and V. Schachter, "Iterative reconstruction of a global metabolic model of Acinetobacter baylyi ADP1 using high-throughput growth phenotype and gene essentiality data," *BMC Syst. Biol.*, vol. 2, 2008.

[44] M. A. Islam, K. Zengler, E. A. Edwards, R. Mahadevan, and G. Stephanopoulos, "Investigating Moorella thermoacetica metabolism with a genome-scale constraint-based metabolic model," *Integr. Biol.*, vol. 7, no. 8, pp. 869–882, 2015.

[45] B. Teusink, A. Wiersma, D. Molenaar, C. Francke, W. M. De Vos, R. J. Siezen, and E. J. Smid, "Analysis of growth of Lactobacillus plantarum WCFS1 on a complex medium using a genome-scale metabolic model," *J. Biol. Chem.*, vol. 281, no. 52, pp. 40041–40048, 2006.

[46] T. T. Vu, S. M. Stolyar, G. E. Pinchuk, E. A. Hill, L. A. Kucek, R. N. Brown, M. S. Lipton, A. Osterman, J. K. Fredrickson, A. E. Konopka, A. S. Beliaev, and J. L. Reed, "Genome-scale modeling of light-driven reductant partitioning and carbon fluxes in diazotrophic unicellular cyanobacterium Cyanothece sp. ATCC 51142," *PLoS Comput. Biol.*, vol. 8, no. 4, 2012.

[47] C. M. aria González-Domenech, E. Belda, R. Patiño-Navarrete, A. Moya, J. Peretó, and A. Latorre, "Metabolic stasis in an ancient symbiosis: genome-scale metabolic networks from two Blattabacterium cuenoti strains, primary endosymbionts of cockroaches," *BMC Microbiol.*, vol. 12, no. Suppl 1, p. S5, 2012.

[48] C. B. Milne, J. A. Eddy, R. Raju, S. Ardekani, P. J. Kim, R. S. Senger, Y. S. Jin, H. P. Blaschek, and N. D. Price, "Metabolic network reconstruction and genome-scale model of butanol-producing strain Clostridium beijerinckii NCIMB 8052," *BMC Syst. Biol.*, vol. 5, 2011.

[49] C. Risso, J. Sun, K. Zhuang, R. Mahadevan, R. DeBoy, W. Ismail, S. Shrivastava, H. Huot, S. Kothari, S. Daugherty, O. Bui, C. H. Schilling, D. R. Lovley, and B. A. Methé, "Genome-scale comparison and constraint-based metabolic reconstruction of the facultative anaerobic Fe(III)-reducer Rhodoferax ferrireducens," *BMC Genomics*, vol. 10, p. 447, 2009.

[50] T. J. Mueller, B. M. Berla, H. B. Pakrasi, and C. D. Maranas, "Rapid construction of metabolic models for a family of Cyanobacteria using a multiple source annotation workflow," *BMC Syst. Biol.*, vol. 7, pp. 1–12, 2013.

[51] J. T. Broddrick, B. E. Rubin, D. G. Welkie, N. Du, N. Mih, S. Diamond, J. J. Lee, S. S. Golden, and B. O. Palsson, "Unique attributes of cyanobacterial metabolism revealed by improved genome-scale metabolic modeling and essential gene analysis," *Proc. Natl. Acad. Sci.*, vol. 113, no. 51, pp. E8344–E8353, 2016.

[52] J. Lee, H. Yun, A. M. Feist, B. Palsson, and S. Y. Lee, "Genome-scale reconstruction and in silico analysis of the Clostridium acetobutylicum ATCC 824 metabolic network," *Appl. Microbiol. Biotechnol.*, vol. 80, no. 5, pp. 849–862, 2008.

[53] J. Levering, T. Fiedler, A. Sieg, K. W. A. van Grinsven, S. Hering, N. Veith, B. G. Olivier, L. Klett, J. Hugenholtz, B. Teusink, B. Kreikemeyer, and U. Kummer, "Genome-scale reconstruction of the Streptococcus pyogenes M49 metabolic network reveals growth requirements and indicates potential drug targets," *J. Biotechnol.*, vol. 232, pp. 25–37, 2016.

[54] J. Sun, B. Sayyar, J. E. Butler, P. Pharkya, T. R. Fahland, I. Famili, C. H. Schilling, D. R. Lovley, and R. Mahadevan, "Genome-scale constraint-based modeling of Geobacter metallireducens," *BMC Syst. Biol.*, vol. 3, pp. 1–15, 2009.

[55] M. Fondi, I. Maida, E. Perrin, A. Mellera, S. Mocali, E. Parrilli, M. L. Tutino, P. Liò, and R. Fani, "Genome scale metabolic reconstruction and constraints-based modelling of the Antarctic bacterium Pseudoalteromonas haloplanktis TAC125," *Environ. Microbiol.*, p. n/a-n/a, 2014.

[56] M. C. Gonnerman, M. N. Benedict, A. M. Feist, W. W. Metcalf, and N. D. Price, "Genomically and biochemically accurate metabolic reconstruction of Methanosarcina barkeri Fusaro, iMG746," *Biotechnol. J.*, vol. 8, no. 9, pp. 1070–1079, 2013.

[57] M. Ponce-de-León, F. Montero, and J. Peretó, "Solving gap metabolites and blocked reactions in genome-scale models: Application to the metabolic network of Blattabacterium cuenoti," *BMC Syst. Biol.*, vol. 7, 2013.

[58] N. A. L. Flahaut, A. Wiersma, B. Van De Bunt, D. E. Martens, P. J. Schaap, L. Sijtsma, V. A. M. Dos Santos, and W. M. De Vos, "Genome-scale metabolic model for Lactococcus lactis MG1363 and its application to the analysis of flavor formation," *Appl. Microbiol. Biotechnol.*, vol. 97, no. 19, pp. 8729–8739, 2013.

[59] P. F. Suthers, M. S. Dasika, V. S. Kumar, G. Denisov, J. I. Glass, and C. D. Maranas, “Genome-scale metabolic reconstruction Of mycoplasma genitalium, iPS189,” *PLoS Comput. Biol.*, vol. 5, no. 2, 2009.

[60] A. Raghunathan, J. Reed, S. Shin, B. Palsson, and S. Daefler, “Constraint-based analysis of metabolic capacity of Salmonella typhimurium during host-pathogen interaction,” *BMC Syst. Biol.*, vol. 3, pp. 1–16, 2009.

[61] W. Zou, L. Liu, J. Zhang, H. Yang, M. Zhou, Q. Hua, and J. Chen, “Reconstruction and analysis of a genome-scale metabolic model of the vitamin C producing industrial strain Ketogulonicigenium vulgare WSH-001,” *J. Biotechnol.*, vol. 161, no. 1, pp. 42–48, 2012.

[62] T. A. Mendum, J. Newcombe, A. A. Mannan, A. M. Kierzek, and J. McFadden, “Interrogation of global mutagenesis data with a genome scale model of Neisseria meningitidis to assess gene fitness in vitro and in sera,” *Genome Biol.*, vol. 12, no. 12, 2011.

[63] M. J. McAnulty, J. Y. Yen, B. G. Freedman, and R. S. Senger, “Genome-scale modeling using flux ratio constraints to enable metabolic engineering of clostridial metabolism in silico,” *BMC Syst. Biol.*, vol. 6, 2012.

[64] S. Dash, T. J. Mueller, K. P. Venkataramanan, E. T. Papoutsakis, and C. D. Maranas, “Capturing the response of Clostridium acetobutylicum to chemical stressors using a regulated genome-scale metabolic model,” *Biotechnol. Biofuels*, vol. 7, no. 1, pp. 1–16, 2014.

[65] R. Saha, A. T. Verseput, B. M. Berla, T. J. Mueller, H. B. Pakrasi, and C. D. Maranas, “Reconstruction and Comparison of the Metabolic Potential of Cyanobacteria Cyanothece sp. ATCC 51142 and Synechocystis sp. PCC 6803,” *PLoS One*, vol. 7, no. 10, 2012.

[66] E. Motamedian, M. Saeidi, and S. A. Shojaosadati, “Reconstruction of a charge balanced genome-scale metabolic model to study the energy-uncoupled growth of Zymomonas mobilis ZM1,” *Mol. BioSyst.*, vol. 12, no. 4, pp. 1241–1249, 2016.

[67] J. I. Hendry, C. B. Prasannan, A. Joshi, S. Dasgupta, and P. P. Wangikar, “Metabolic model of Synechococcus sp. PCC 7002: Prediction of flux distribution and network modification for enhanced biofuel production,” *Bioresour. Technol.*, vol. 213, pp. 190–197, 2016.

[68] J. Puchałka, M. A. Oberhardt, M. Godinho, A. Bielecka, D. Regenhardt, K. N. Timmis, J. A. Papin, and V. A. P. Martins Dos Santos, "Genome-scale reconstruction and analysis of the Pseudomonas putida KT2440 metabolic network facilitates applications in biotechnology," *PLoS Comput. Biol.*, vol. 4, no. 10, 2008.

[69] M. a. Oberhardt, J. Puchałka, V. a P. M. dos Santos, and J. a. Papin, "Reconciliation of genome-scale metabolic reconstructions for comparative systems analysis," *PLoS Comput. Biol.*, vol. 7, no. 3, 2011.

[70] P. Babaei, S.-A. Marashi, and S. Asad, "Genome-scale reconstruction of the metabolic network in Pseudomonas stutzeri A1501," *Mol. BioSyst.*, vol. 11, no. 11, pp. 3022–3032, 2015.

[71] G. E. Pinchuk, E. A. Hill, O. V. Geydebrekht, J. de Ingeniis, X. Zhang, A. Osterman, J. H. Scott, S. B. Reed, M. F. Romine, A. E. Konopka, A. S. Beliaev, J. K. Fredrickson, and J. L. Reed, "Constraint-based model of Shewanella oneidensis MR-1 metabolism: A tool for data analysis and hypothesis generation," *PLoS Comput. Biol.*, vol. 6, no. 6, pp. 1–8, 2010.

[72] I. Borodina, P. Krabben, and J. Nielsen, "Genome-scale analysis of Streptomyces coelicolor A3(2) metabolism," *Genome Res.*, vol. 3, no. 2, pp. 820–829, 2005.

[73] M. W. Kim, J. S. Yi, J. J. Kim, J. J. Kim, M. W. Kim, G. Kim, B. Engineering, and S. National, "Reconstruction of high-quality metabolic model enables identification of gene overexpression targets for enhanced antibiotics production in Streptomyces coelicolor A3(2)," *Biotechnol. J.*, pp. 1–33, 2014.

[74] M. A. Oberhardt, J. Puchałka, K. E. Fryer, V. A. P. Martins Dos Santos, and J. A. Papin, "Genome-scale metabolic network analysis of the opportunistic pathogen Pseudomonas aeruginosa PAO1," *J. Bacteriol.*, vol. 190, no. 8, pp. 2790–2803, 2008.

[75] M. I. Pastink, B. Teusink, P. Hols, S. Visser, W. M. De Vos, and J. Hugenholtz, "Genome-scale model of Streptococcus thermophilus LMG18311 for metabolic comparison of lactic acid bacteria," *Appl. Environ. Microbiol.*, vol. 75, no. 11, pp. 3627–3633, 2009.

[76] W. Zou, M. Zhou, L. Liu, and J. Chen, "Reconstruction and analysis of the industrial strain Bacillus megaterium WSH002 genome-scale in silico metabolic model," *J. Biotechnol.*, vol. 164, no. 4, pp. 503–509, 2013.

[77] N. Veith, M. Solheim, K. W. A. van Grinsven, B. G. Olivier, J. Levering, R. Grosseholz, J. Hugenholtz, H. Holo, I. Nes, B. Teusink, and U. Kummer, "Using a genome-scale metabolic model of Enterococcus faecalis V583 to assess amino acid uptake and its impact on central metabolism," *Appl. Environ. Microbiol.*, vol. 81, no. 5, pp. 1622–1633, 2015.

[78] R. Mahadevan, D. R. Bond, J. E. Butler, V. Coppi, B. O. Palsson, C. H. Schilling, and D. R. Lovley, "Characterization of Metabolism in the Fe (III) -Reducing Organism Geobacter sulfurreducens by Constraint-Based Modeling Characterization of Metabolism in the Fe (III) -Reducing Organism Geobacter sulfurreducens by Constraint-Based Modeling †," *Appl. Environ. Microbiol.*, vol. 72, no. 2, pp. 1558–1568, 2006.

[79] Saheed Imam, S. Yilmaz, U. Sohmen, A. S. Gorzalski, J. L. Reed, D. R. Noguera, and T. J. Donohue, "iRsp1095: A genome-scale reconstruction of the Rhodobacter sphaeroides metabolic network," *BMC Syst. Biol.*, vol. 5, no. 116, pp. 1–16, 2011.

[80] S. Imam, D. R. Noguera, and T. J. Donohue, "Global insights into energetic and metabolic networks in Rhodobacter sphaeroides," *BMC Syst. Biol.*, vol. 7, no. 89, 2013.

[81] N. Xu, J. Liu, L. Ai, and L. Liu, "Reconstruction and analysis of the genome-scale metabolic model of Lactobacillus casei LC2W," *Gene*, vol. 554, no. 2, pp. 140–147, 2015.

[82] N. Swainston, K. Smallbone, P. Mendes, D. Kell, and N. Paton, "The SuBliMinaL Toolbox: automating steps in the reconstruction of metabolic networks.," *J. Integr. Bioinform.*, vol. 8, no. 2, p. 186, 2011.

[83] Y. Wang, N. Xu, C. Ye, L. Liu, Z. Shi, and J. Wu, "Reconstruction and in silico analysis of an Actinoplanes sp. SE50/110 genome-scale metabolic model for acarbose production," *Front. Microbiol.*, vol. 6, no. JUN, pp. 1–14, 2015.

[84] K. Yoshikawa, S. Aikawa, Y. Kojima, Y. Toya, C. Furusawa, A. Kondo, and H. Shimizu, "Construction of a genome-scale metabolic model of arthrospira platensis nies-39 and metabolic design for cyanobacterial bioproduction," *PLoS One*, vol. 10, no. 12, pp. 1–16, 2015.

[85] A. Metris, M. Reuter, D. J. H. Gaskin, J. Baranyi, and A. H. M. van Vliet, "In vivo and in silico determination of essential genes of Campylobacter jejuni," *BMC Genomics*, vol. 12, 2011.

[86] M. A. Richards, T. J. Lie, J. Zhang, S. W. Ragsdale, J. A. Leigh, and N. D. Price, “Exploring hydrogenotrophic methanogenesis: A genome scale metabolic reconstruction of Methanococcus maripaludis,” *J. Bacteriol.*, vol. 198, no. 24, pp. 3379–3390, 2016.

[87] Y. Shinfuku, N. Sorpitiporn, M. Sono, C. Furusawa, T. Hirasawa, and H. Shimizu, “Development and experimental verification of a genome-scale metabolic model for Corynebacterium glutamicum,” *Microb. Cell Fact.*, vol. 8, pp. 1–15, 2009.

[88] M. Latendresse, M. Krummenacker, M. Trupp, and P. D. Karp, “Construction and completion of flux balance models from pathway databases,” *Bioinformatics*, vol. 28, no. 3, pp. 388–396, 2012.

[89] V. Satish Kumar, M. S. Dasika, and C. D. Maranas, “Optimization based automated curation of metabolic reconstructions,” *BMC Bioinformatics*, vol. 8, pp. 1–16, 2007.

[90] J. D. Orth, T. M. Conrad, J. Na, J. a Lerman, H. Nam, A. M. Feist, and B. Ø. Palsson, “A comprehensive genome-scale reconstruction of Escherichia coli metabolism—2011,” *Mol. Syst. Biol.*, vol. 7, no. 535, pp. 1–9, 2011.

[91] A. Bordbar, H. Nagarajan, N. E. Lewis, H. Latif, A. Ebrahim, S. Federowicz, J. Schellenberger, and B. O. Palsson, “Minimal metabolic pathway structure is consistent with associated biomolecular interactions.,” *Mol. Syst. Biol.*, vol. 10, p. 737, 2014.

[92] J. Ramnauth, M. D. Surman, P. B. Sampson, B. Forrest, J. Wilson, E. Freeman, D. D. Manning, F. Martin, A. Toro, M. Domagala, D. E. Awrey, E. Bardouniotis, N. Kaplan, J. Berman, and H. W. Pauls, “Bioorganic & Medicinal Chemistry Letters of the bacterial enoyl ACP reductase , FabI,” *Bioorg. Med. Chem. Lett.*, vol. 19, no. 18, pp. 5359–5362, 2009.

[93] J. Jo and S. M. Raj, “Cloning , expression , and characterization of an aldehyde dehydrogenase from Escherichia coli K-12 that utilizes 3-Hydroxypropionaldehyde as a substrate,” pp. 51–60, 2008.

[94] M. Donnelly and R. Cooper, “Succinic Semialdehyde Dehydrogenases of Escherichia coli Their Role in the Degradation of p-Hydroxyphenylacetate and y-Aminobutyrate,” *Eur. J. Biochem.*, vol. 561, pp. 555–561, 1981.

[95] A. Gruez, S. Grisel, C. Valencia, M. Tegoni, and C. Cambillau, “Crystal Structure and Kinetics

Identify Escherichia coli YdcW Gene Product as a Medium-chain Aldehyde Dehydrogenase," pp. 29–41, 2004.

[96] C. Bausch, N. Peekhaus, C. Utz, T. Blais, E. Murray, T. Lowary, and T. Conway, "Sequence Analysis of the GntII (Subsidiary) System for Gluconate Metabolism Reveals a Novel Pathway for L -Idonic Acid Catabolism in Escherichia coli," vol. 180, no. 14, pp. 3704–3710, 1998.

[97] a M. Gardner, L. a Martin, P. R. Gardner, Y. Dou, and J. S. Olson, "Steady-state and transient kinetics of {E}scherichia coli nitr ic-oxide dioxygenase (flavohemoglobin)," *J. Biol. Chem.*, vol. 275, no. 17, pp. 12581–12589, 2000.

[98] F. Nunez, M. T. Pellicer, J. Badia, J. Aguilar, and L. Baldoma, "Biochemical characterization of the 2-ketoacid reductases encoded by ycdW and yiaE genes in Escherichia coli," *Biochem. J.*, vol. 715, pp. 707–715, 2001.

[99] D. Yum, B. Lee, and D. Hahm, "The yiaE Gene , Located at 80 . 1 Minutes on the Escherichia coli Chromosome , Encodes a 2-Ketoaldonate Reductase The yiaE Gene , Located at 80 . 1 Minutes on the Escherichia coli Chromosome , Encodes a 2-Ketoaldonate Reductase," *J. Bacteriol.*, vol. 180, no. 22, pp. 1–6, 1998.

[100] Y. Nakano, N. Suzuki, Y. Yoshida, T. Nezu, Y. Yamashita, and T. Koga, "Thymidine Diphosphate-6-deoxy- L - lyxo -4-hexulose Reductase Synthesizing dTDP-6-deoxy- L -talose from Actinobacillus actinomycetemcomitans *," vol. 275, no. 10, pp. 6806–6812, 2000.

[101] J. R. Edgar and M. Bell, "Biosynthesis in Escherichia coli of sn-Glycerol-3-Phosphate , a Precursor of Phospholipid," *J. Biol. Chem.*, vol. 255, no. 8, pp. 3492–3497, 1979.

[102] T. Lacour, T. Achstetter, and B. Dumas, "Characterization of recombinant adrenodoxin reductase homologue (Arh1p) from yeast. Implication in in vitro cytochrome P45011β monooxygenase system," *J. Biol. Chem.*, vol. 273, no. 37, pp. 23984–23992, 1998.

[103] X. Wang, C. J. Mann, Y. Bai, L. Ni, and H. Weiner, "Molecular cloning, characterization, and potential roles of cytosolic and mitochondrial aldehyde dehydrogenases in ethanol metabolism in Saccharomyces cerevisiae.," *J. Bacteriol.*, vol. 180, no. 4, pp. 822–830, 1998.

[104] F. RAMOS, M. El GUEZZAR, M. GRENSON, and J. -M WIAME, "Mutations affecting the

enzymes involved in the utilization of 4-aminobutyric acid as nitrogen source by the yeast Saccharomyces cerevisiae," *Eur. J. Biochem.*, vol. 149, no. 2, pp. 401–404, 1985.

[105] B. DeLaBarre, P. R. Thompson, G. D. Wright, and A. M. Berghuis, "Crystal structures of homoserine dehydrogenase suggest a novel catalytic mechanism for oxidoreductases," *Nat. Struct. Biol.*, vol. 7, no. 3, pp. 238–249, 2000.

[106] K. Baudry, E. Swain, A. Rahier, M. Germann, A. Batta, S. Rondet, S. Mandala, K. Henry, G. S. Tint, T. Edlind, M. Kurtz, and J. T. Nickels, "The Effect of the *erg26-1* Mutation on the Regulation of Lipid Metabolism in *Saccharomyces cerevisiae*," *J. Biol. Chem.*, vol. 276, no. 16, pp. 12702–12711, 2001.

[107] J. Forster, I. Famili, B. O. Palsson, and J. Nielsen, "Genome-scale reconstruction of the *Saccharomyces cerevisiae* metabolic network," *Genome Res.*, no. 13, pp. 244–253, 2003.

[108] M. K. Jensen and J. D. Keasling, "Recent applications of synthetic biology tools for yeast metabolic engineering," *FEMS Yeast Res.*, vol. 15, no. 1, pp. 1–10, 2015.

[109] T. Österlund, I. Nookaew, and J. Nielsen, "Fifteen years of large scale metabolic modeling of yeast: Developments and impacts," *Biotechnol. Adv.*, vol. 30, no. 5, pp. 979–988, 2012.

[110] M. J. Herrgård, N. Swainston, P. Dobson, W. B. Dunn, K. Y. Arga, M. Arvas, N. Blüthgen, S. Borger, R. Costenoble, M. Heinemann, M. Hucka, N. Le Novère, P. Li, W. Liebermeister, M. L. Mo, A. P. Oliveira, D. Petranovic, S. Pettifer, E. Simeonidis, K. Smallbone, I. Spasi, D. Weichart, R. Brent, D. S. Broomhead, H. V Westerhoff, B. Kırdar, M. Penttilä, E. Klipp, B. Ø. Palsson, U. Sauer, S. G. Oliver, P. Mendes, J. Nielsen, and D. B. Kell, "A consensus yeast metabolic network reconstruction obtained from a community approach to systems biology," *Nat. Biotechnol.*, vol. 26, no. 10, pp. 1155–1160, 2008.

[111] N. C. Duarte, M. J. Herrgard, and B. O. Palsson, "Reconstruction and Validation of Saccharomyces cerevisiae iND750, a Fully Compartmentalized Genome-Scale," *Genome Res.*, no. 14, pp. 1298–1309, 2004.

[112] L. Kuepfer, U. Sauer, and L. M. Blank, "Metabolic functions of duplicate genes in Saccharomyces cerevisiae," *Genom*, vol. 15, pp. 1421–1430, 2005.

[113] T. A. Pemberton, D. Srivastava, N. Sanyal, M. T. Henzl, D. F. Becker, and J. J. Tanner,

"Structural studies of yeast Δ1-pyrroline-5-carboxylate dehydrogenase (ALDH4A1): Active site flexibility and oligomeric state," *Biochemistry*, vol. 53, no. 8, pp. 1350–1359, 2014.

[114] D. W. Lundgren and M. Ogur, "INHIBITION OF YEAST A1-PYRROLINE-5-CARBOXYLATE DEHYDROGENASE BY COMMON AMINO ACIDS AND THE REGULATION OF PROLINE CATABOLISM," *Biochim. Biophys. Acta*, vol. 297, pp. 246–257, 1973.

[115] J. P. van Dijken, R. A. Weusthuis, and J. T. Pronk, "Kinetics of growth and sugar consumption in yeasts," *Antonie Van Leeuwenhoek*, vol. 63, no. 3–4, pp. 343–352, 1993.

[116] R. Pereira, J. Nielsen, and I. Rocha, "Improving the flux distributions simulated with genome-scale metabolic models of Saccharomyces cerevisiae," *Metab. Eng. Commun.*, vol. 3, pp. 153–163, 2016.

[117] T. M. Loftus, L. V. Hall, S. L. Anderson, and L. McAlister-Henn, "Isolation, Characterization, and Disruption of the Yeast Gene Encoding Cytosolic NADP-Specific Isocitrate Dehydrogenase," *Biochemistry*, vol. 33, no. 32, pp. 9661–9667, 1994.

[118] R. J. Haselbeck and L. Mcalister-henns, "Function and Expression of Yeast Mitochondrial NAD- and NADP-specific Isocitrate Dehydrogenases *," *J. Biol. Chem.*, vol. 268, no. 16, pp. 12116–12122, 1993.

[119] J. Satrustegui, J. Bautista, and A. Machado, "NADPH/NADP+ ratio: regulatory implications in yeast glyoxylic acid cycle," *Mol. Cell. Biochem.*, vol. 51, no. 2, pp. 123–127, 1983.

[120] A. K. Gombert and M. Moreira, "Network Identification and Flux Quantification in the Central Metabolism of *Saccharomyces cerevisiae* under Different Conditions of Glucose Repression Network Identification and Flux Quantification in the Central Metabolism of Saccharomyces cerev," *J. Bacteriol.*, vol. 183, no. 4, pp. 1441–1451, 2001.

[121] L. Jahnke and H. P. Klein, "Oxygen requirements for formation and activity of the squalene epoxidase in Saccharomyces cerevisiae.," *J. Bacteriol.*, vol. 155, no. 2, pp. 488–492, 1983.

CHAPTER 5

Development of an *in silico* method for the conversion of NAD(P)(H) cofactor specificity in structurally uncharacterized enzymes

The information presented in this Chapter has been submitted for a patent application:

Resende T., Soares C., Rocha I. A method for the *in silico* conversion of NAD(P)(H) cofactor specificity in structurally uncharacterized enzymes

5.1 Introduction

Since the establishment of the industrial revolution, fossil fuels such as crude, gas and coal, have played a decisive role in the rapid development of several industries responsible for the steady growth of world economies. However, as climate changes are becoming a real threat, and more prominently due to the rising fear shared by world leading economies concerning natural or human imposed fossil fuel scarcity, this reality is transforming.

For many years now, the world has been making a huge effort in enhancing its fossil fuel supported economy with a safer, healthier and environmentally conscious biotechnological approach. These efforts supported the growth of multiple new biotechnology fields such as synthetic biology and metabolic engineering, with a clear focus on the development of biosynthetic chemicals, pharmaceutical drugs and in the production of efficient biocatalysts for replacing chemically driven reactions [1]–[4].

Despite great advances and breakthroughs achieved by the aforementioned biotechnological fields, the complex metabolic panorama of most organisms is still elusive, hindering an efficient and replicable tuning of their biosynthetic pathways for the optimization of desired by-products [5].

One of the most relevant challenges in designing biosynthetic pathways in an organism's metabolism resides in the fragile balancing of reaction cofactors availability, such as the case of nicotinamide adenine dinucleotide (NAD(H)) and nicotinamide adenine dinucleotide phosphate (NADP(H)) [6]. Being the most widely used cofactors, cell metabolism heavily regulates the levels of reduced and oxidized metabolic pools of NAD(P)(H), which are often targets for the metabolic engineering of biological pathways and systems [7]. The fine balance between these seemingly equivalent cofactors has been the subject of multiple studies [8]–[10], with several approaches being undertaken in order to modify the related pathways in the process of engineering biosynthetic production [11]–[14]. Cyanobacteria, as an example, have been genetically modified to produce tangible amounts of added-value biochemicals, such us biofuels and lactate. However, their metabolic unbalance towards NADP(H) limits the synthesis of these products. Approaches for solving this problem include the usage of heterologous enzymes with a different cofactor usage, the over expression of cofactor synthesis and also metabolic engineering for cofactor specificity reversal [8].

This problem has also been addressed, *in silico*, using genome-scale metabolic models (GEMs) with approaches being developed in an attempt of simplifying and optimizing cofactor balancing for strain design endeavors [15], [16]. The computational method *OptSwap* encompasses an optimization protocol developed specifically to address this problem. Using GEMs, this method predicts strain designs by identifying optimal modifications of NAD(P)(H) cofactor binding enzymes as well as complementary reaction knockouts [16]. However, despite its relevance and applicability, the mentioned method refrains to refer any approach on how to accomplish such cofactor binding alteration.

Due to the structural similarity between NAD(H) and NADP(H), with their only difference residing in the presence of a phosphate group in the vicinity of the 2' hydroxyl of the adenosine ribose in NADP(H), specificity mechanisms are difficult to characterize, hindering rational approaches for performing NAD(P)(H) cofactor specificity reversal. Nonetheless, reversing NAD(P)(H) cofactor specificity is a frequently addressed problem in metabolic engineering and strain design due to its implications in metabolic pathway flux redirection. Cui *et al*, [17] compiled a list of enzyme redesigns for altered NAD(P)(H) cofactor specificity using site-directed mutagenesis, showing that multiple simultaneous mutations have to be performed in order to effectively change cofactor specificity, due to their non-additive characteristics.

In addition to their unpredictability, current methodologies for performing NAD(P)(H) cofactor specificity change require intensive experimental work, being performed primarily on a case-by-case workflow, with gained knowledge from previous experiments being often not able to be reproduced in new ones.

One of the latest methods developed to aid in this arduous task works by targeting the residues contacting the 2'moiety directly, or through water mediated interactions, with a designed library of mutations[18]. Target residues are classified according to their relative position towards the nicotinamide moiety [19] and mutations are designed utilizing subsaturation degenerate codon libraries. However, this approach is subjective as library design is based in the inclusion of mutations previously described in the literature as valuable for cofactor specificity reversal [18]. Moreover, the amount of experimental work required for processing the designed libraries is still enormous and, also, a characterized enzyme's structure, with the target cofactor bound, as well as information on cofactor specificity, are still required.

In chapter 3, we set out to unveil the molecular determinants for NAD(P)(H) cofactor specificity, using enzyme structural information. Using support vector machines (SMV) we were able to identify and categorize, by cofactor specificity, aminoacid residues positioned around different areas of the binding of cofactors in a large enzyme dataset. With the development of a specificity prediction tool, these findings were successfully applied in the prediction of cofactor specificity of enzymes not structurally characterized.

In the present chapter we propose a new method for the *in silico* efficient conversion of NAD(P)(H) cofactor specificity in non structurally characterized enzymes.

Firstly, a method for identifying and producing an ordered list containing the most influential residues for cofactor specificity in a given enzyme structure was developed, using the SVM predictive model and CNRPM (Cofactor Neighbor Residue Profile Matrix) protocol generated in chapter 3. The created list is assumed as containing the theoretical optimal set of residue positions suitable for point mutations conferring cofactor specificity reversal.

Secondly, two methods were developed in order to identify the optimal set of point mutations required for achieving cofactor specificity change. For that, two distinct approaches were implemented, being one deterministic, using the gathered information on the most influential residues for both cofactor specificities; and the other stochastic, using evolutionary algorithms to locate the optimal set of mutations capable of reverting cofactor specificity.

With the implementation of the developed methods, we were able to pinpoint theoretical optimal aminoacid residue mutations for an efficient alteration of cofactor specificity in case studies, and predict the resulting cofactor specificity using the developed tool for cofactor prediction.

5.2 Methods

5.2.1 Aminoacid residue sequences and protein structure templates

The wild-type amino acid residue sequences from the selected case-study enzymes were retrieved from Uniprot, while mutant sequences were replicated *in silico* according to the specifications presented in the literature. Uniprot IDs from the four selected enzymes are as following: *Pichia stipitis*'s xylose reductase: P31867; *Gluconobacter oxydans*'s xylitol dehydrogenase: Q8GR61; *Pichia stipitis*'s xylitol dehydrogenase: P22144; *Tramitichromis intermedius*'s leucine dehydrogenase: Q60030.

Gluconobacter oxydans' xylitol dehydrogenase had its structure already experimentally characterized (PDB id: 1ZEM), being therefore used as a template for modeling the structure of the predicted mutants. As for the remaining three enzymes, since their structures were not experimentally characterized, the structures of the wild-type and corresponding mutants were generated using comparative modeling. *Pichia stipitis'* xylose reductase structure was modeled using the homologue structure of *Arabidopsis thaliana*'s aldo-keto reductase (PDB id: 3H7R) with 47% identity; *Pichia stipitis*' xylitol dehydrogenase structure was modeled using the homologue structure of *Homo sapiens*' sorbitol dehydrogenase (PDB id: 1PL6) with 44% identity and *Tramitichromis intermedius*' leucine dehydrogenase structure was modeled using the homologue structure of *Rhodococcus sp*. M4's phenylalanine dehydrogenase (PDB id: 1BW9) with 37% identity.

5.2.2 Cofactor specificity prediction

Predictions on cofactor specificity change were performed using NiCofactor, the developed tool described in chapter 3, for allowing the high throughput NAD(P)(H) cofactor specificity prediction. NiCofactor was built using the python programming language. Input sequences are required to be in FASTA format. For each sequence, the tool initiates an individual project. The tools for generating CNRPMs and performing machine learning were also integrated in NiCofactor. Results are outputted by attributing to each analyzed sequence a cofactor prediction and subsequent prediction score. The default probability score threshold used is 0.8.

5.2.3 Evolutionary algorithm implementation

The evolutionary algorithms used in the stochastic method for efficiently predict the optimal set of mutations to reverse cofactor specificity were implemented using *inspyred* [20], an open source framework for creating biologically-inspired computational intelligence algorithms in Python.

Five evolutionary algorithms were implemented, with the only difference between them being the maximum candidate size allowed. Each algorithm was configured to run through 100 generations, being the initial population composed by 100 individuals. Each individual was randomly created according to the candidate maximum size, which varied between 1 and 5. This corresponds to the creation of mutant aminoacid residue sequences derived from the original target aminoacid residue sequence, containing between 1 and 5 mutations each. Elitism value was set to 2, keeping the best 2 scoring individuals for the next generation. The next best scoring 50 individuals were recombined using mutation operators, with a crossover rate of 0.9 and a mutation rate of 0.1. The crossover operator uses the parameters of two individuals and combines them, generating two new individuals, while the mutation operator substitutes one element of the individual by another, randomly generated. The remaining lowest scoring 48 individuals were discarded and newly generated individuals with random mutations in the available mutable positions were incorporated in the population. The optimization process is terminated when the maximum number of generations is achieved.

5.2.4 Protein structure visualization

Wild-type and mutant enzyme structures were visualized using PyMol [21] a free and user-friendly molecular graphics system for molecular visualization written in Python programming language.

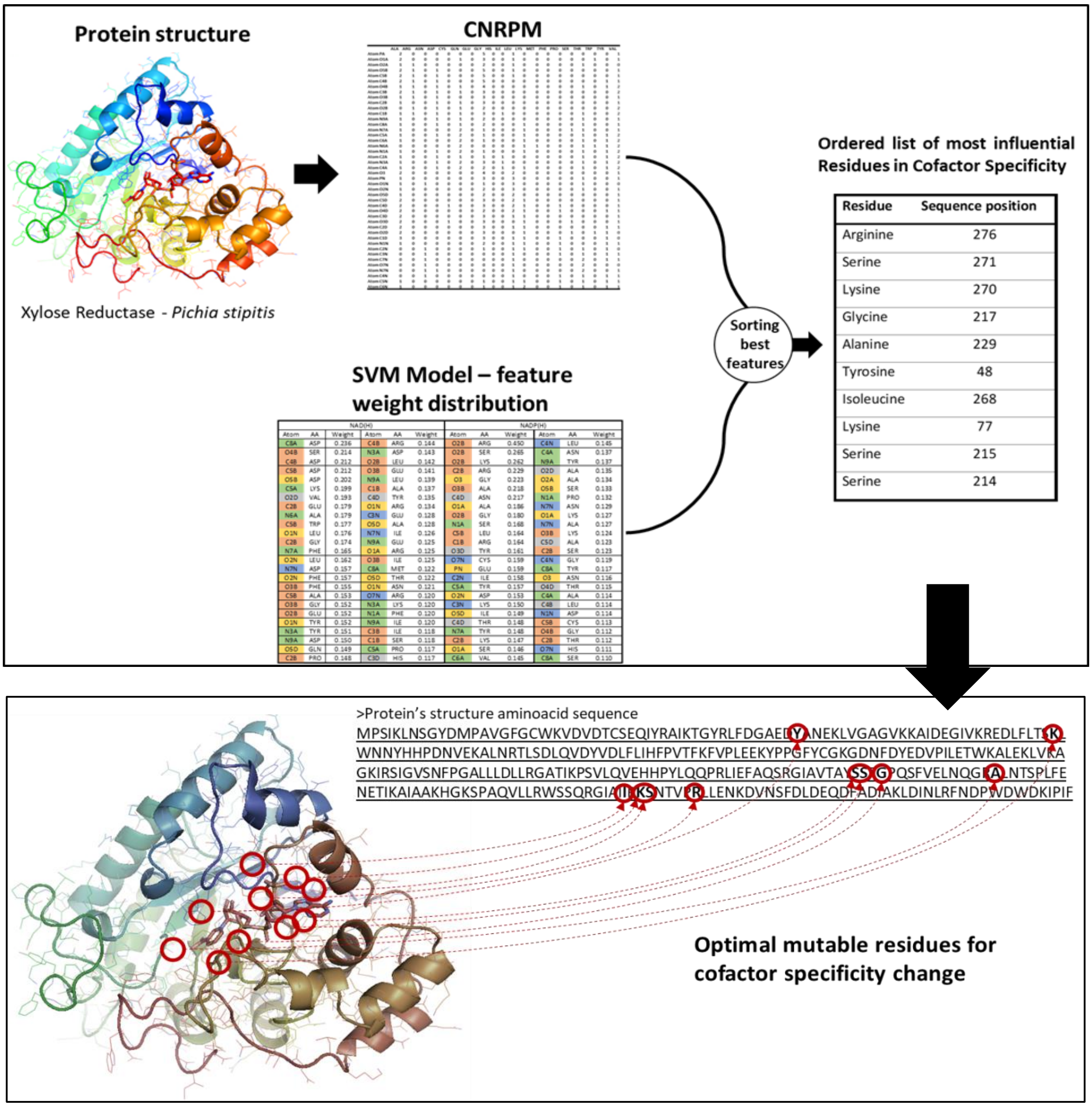
5.3 Results and discussion

5.3.1 Identification of mutable residue positions for cofactor specificity reversal

As previously stated in chapter 3, the SVM model, trained with a large dataset of CNRPMs, performed the attribution of importance scores to the features in the dataset, allowing the correct separation of the instances in the hyperplane. In this case, the features are composed of relations between the atoms in each cofactor and the corresponding neighbor aminoacid residues. Then, the score attributed to each feature, reveals the impact of each interaction in the binding preference of the cofactors, with higher scoring features having a higher impact on cofactor specificity. Figure 5.1 depicts a representation of the process undertaken for the selection of mutable residue positions for cofactor specificity reversal. Through the analysis of the CNRPM generated from a target enzyme's structure, and by combining this information with the data stored in the SVM model, the sorting of the best features for cofactor specificity is performed. When the most influential features for cofactor specificity in a given protein structure are sorted, the corresponding residue position in the sequence is retrieved and stored for each feature. When features from ten distinct residues in the sequence are retrieved, the list is closed.

The end result is an ordered list of the most influential residues in cofactor specificity for a given target enzyme, being assumed that, due to their influence in cofactor specificity, this list contains the optimal mutable residues for cofactor specificity reversal.

Despite the developed method's utility in precisely pinpointing suitable target residues for cofactor specificity reversal mutations, such endeavor is still greatly hindered by the combinatory amount of total mutant possibilities. In order to optimize this task, two distinct approaches were undertaken, resulting in the development of two different methods for determining the optimal set of mutations necessary to achieve cofactor specificity reversal.


Protein structure
Xylose Reductase - *Pichia stipitis*
CNRPM
SVM Model – feature weight distribution
NAD(H)
NADP(H)
Sorting best features
Ordered list of most influential Residues in Cofactor Specificity
>Protein's structure aminoacid sequence
MPSIKLNSGYDMPAVGFGCWKVDVDTCSEQIYRAIKTGYRLFDGAEDYANEKLVGAGVKKAIDEGIVKREDLFLTSKL
WNNYHHPDNVEKALNRTLSDLQVDYVDLFLIHFPVTFKFVPLEEKYPPGFYCGKGDNFDYEDVPILETWKALEKLVKA
GKIRSIGVSNFPGALLLDLLRGATIKPSVLQVEHHPYLQQPRLIEFAQSRGIAVTAYSSFGPQSFVELNQGRALNTSPLFE
NETIKAIAAKHGKSPAQVLLRWSSQRGIAIIPKSNTVPRLLENKDVNSFDLDEQDFADIAKLDINLRFNDPWDWDKIPIF
Optimal mutable residues for cofactor specificity change

Residue	Sequence position
Arginine	276
Serine	271
Lysine	270
Glycine	217
Alanine	229
Tyrosine	48
Isoleucine	268
Lysine	77
Serine	215
Serine	214

Figure 5.1 – Process for the identification and selection of suitable mutable residue positions for cofactor specificity reversal. By combining the information on the target enzyme's CNRPM with the data in the SVM model, the sorting of best features for a specific enzyme is possible. By returning the features residue position in the sequence, a list of optimal mutable residues for cofactor specificity change is able to be generated.

5.3.2 Cofactor specificity reversal – Deterministic method

With the intent of surpassing the overwhelming amount of combinatorial mutations necessary for the screening of all suitable mutable residue positions to achieve the optimal cofactor specificity reversal mutant, a deterministic method was developed. This deterministic approach is based on the formulation of a hypothesis regarding both cofactor's most influential features. With this in mind, the suggested hypothesis states that: if, for each atom composing a cofactor, there is a specific neighbor aminoacid residue with the highest impact in the cofactor specificity, then, if this interacting aminoacid residue is replaced by the aminoacid residue with the highest impact for the opposite cofactor, the binding specificity should be affected. Therefore, if enough aminoacid residues with high impact on cofactor specificity are changed, the cofactor specificity should be reverted.

In order to implement the stated hypothesis, the extracted SVM feature weights present in appendix, in table A1, were further examined. For each cofactor atom, the strongest feature present for NAD(H) and NADP(H) specificity was selected, and the associated residue retrieved. The resulting chart, displayed in table 5.1, represents, for each cofactor atom, the aminoacid residue interaction with highest impact on cofactor specificity, and consequently the candidate for performing point-mutations in that area of the binding spot. Figure 5.2 depicts the information present in table 5.1 by positioning each aminoacid residue according to the interaction with highest impact on cofactor specificity.

Table 5.1. Point-mutation selecting chart. Each cofactor atom and corresponding molecular localization is represented in the column "cofactor atom". For each atom, the corresponding NAD(H) and NADP(H) specificity is represented in columns "NAD(H) specific" and "NADP(H) specific" respectively.

Cofactor atom	NAD(H) specific	NADP(H) specific	Cofactor atom	NAD(H) specific	NADP(H) specific
PA	GLY	GLU	O3	HIS	GLY
O1A	ARG	ALA	PN	ASN	GLU
O2A	GLU	ALA	O1N	LEU	TRP
O5B	ASP	SER	O2N	LEU	ASP
C5B	ASP	LEU	O5D	GLN	ILE
C4B	ASP	LEU	C5D	SER	ALA
O4B	SER	GLY	C4D	TYR	ASN
C3B	ILE	CYS	O4D	ILE	THR
O3B	PHE	ALA	C3D	HIS	TYR
C2B	GLU	ARG	O3D	THR	TYR
O2B	GLU	ARG	C2D	PRO	GLU
C1B	ALA	ARG	O2D	VAL	ALA
N9A	ASP	TYR	C1D	CYS	THR
C8A	ASP	TYR	N1N	PRO	ASP
N7A	PHE	TYR	C2N	MET	ILE
C5A	LYS	TYR	C3N	GLU	LYS
C6A	ALA	VAL	C7N	ASN	GLN
N6A	ALA	GLN	O7N	ARG	CYS
N1A	PHE	SER	N7N	ASP	ASN
C2A	ASP	ALA	C4N	ALA	LEU
N3A	TYR	ARG	C5N	ASN	ILE
C4A	PHE	ASN	C6N	GLU	PHE

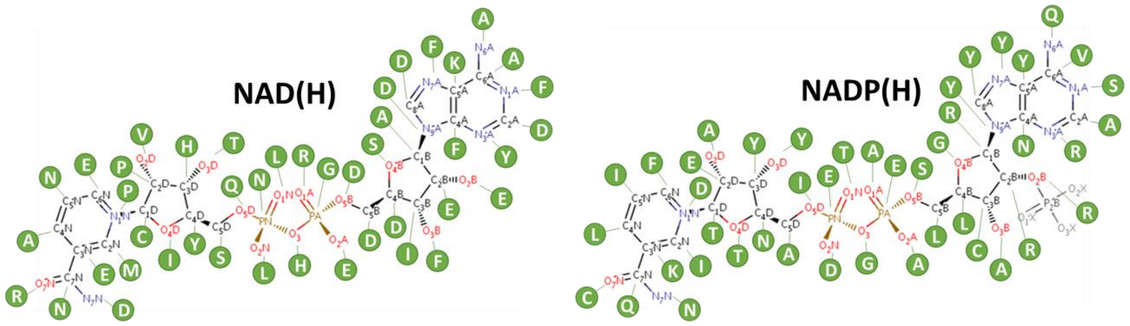


Figure 5.2 - Cofactor atom-aminoacid interactions with highest impact on specificity. Each green circle represents an aminoacid, displayed in the single-letter format, which connects to each corresponding cofactor atom, forming the atom-aminoacid interaction with highest impact on specificity for each atom in both cofactors.

5.3.2.1 Implementation of the deterministic method for cofactor specificity conversion

The implementation of the deterministic method for the conversion of NAD(P)(H) cofactor specificity, depicted in figure 5.3, starts with the structural analysis of the subject enzyme and respectively CNRPM assembly with NiCofactor, the developed tool for predicting cofactor specificity. If the subject enzyme’s structure is not characterized, homology modelling with a suitable structural template is performed automatically by NiCofactor.

Once created, CNRPM features are sorted and the one with the highest impact is selected. The aminoacid residue sequence position from the residue present in the feature is retrieved, while the atom present in the feature is searched in the point-mutation selecting chart displayed in table 5.1, being the candidate aminoacid residue mutant selected.

The wild-type aminoacid residue is replaced by the mutant candidate in the aminoacid residue sequence. The mutant sequence is retrieved and its cofactor specificity is predicted using NiCofactor.

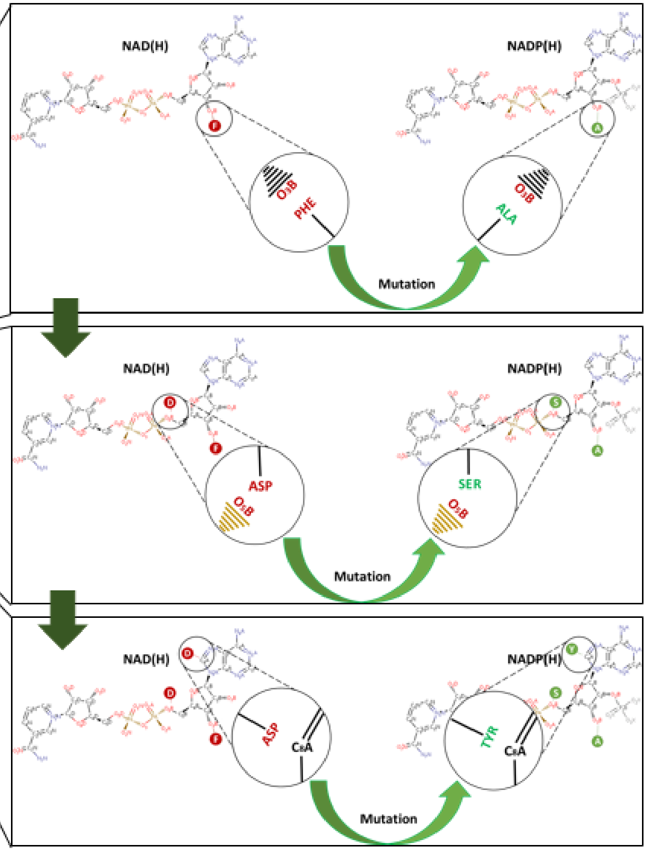
If a prediction is performed, with a probability score above the threshold, and the cofactor predicted for the mutant sequence is changed, the mutant sequence is accepted as having its cofactor specificity successfully altered. If, on the other hand, the cofactor prediction did not change, the conversion method continues with the mutant sequence and the second highest impact feature is used. This step is performed iteratively, with mutations being incremented in the sequence until the cofactor prediction is changed. If after 10 consecutive mutations the cofactor prediction remains unaltered, the mutation is regarded as unviable.

In the example depicted in figure 5.3, for the NAD(H) dependent target enzyme structure, the atom-aminoacid residue interaction with highest impact on cofactor specificity is originated by the presence of a Phenylalanine (F) near the atom O3B, in the ribose from the adenine moiety. By consulting table 5.1 it is possible to observe that the residue originating the atom-aminoacid residue interaction, with atom O3B, with the highest impact on cofactor specificity for NADP(H) is an Alanine (A). Being the sequence position of the selected Phenylalanine, position 42, a point mutation is performed and the Phenylalanine is substituted by an Alanine. Despite this mutation, the cofactor specificity prediction of the target enzyme was not altered, being therefore selected the second highest impact interaction, the Aspartate (D) near the atom O5B. With the substitution of Aspartate by a Serine (S) not rendering an altered cofactor specificity prediction, the third highest impact interaction was selected. This time, the Aspartate near atom C8A was mutated into a Tyrosine (Y) and the resulting mutant F42A/D23S/D71Y was successfully predicted as having reverted its original cofactor specificity.

The presented deterministic method is a fast and precise approach for the complex problem of selecting the optimal set of mutations capable of reverting cofactor specificity in a target enzyme. Despite its overall efficacy, robustness and time efficiency, the deterministic characteristics of this approach mean that there are multiple mutation combinations that are not taken into consideration, with the possibility of better results for a set of cofactor specificity reverting mutations being overlooked. Due to these constraints, and in order to analyze the highest number of mutation combinations possible, a stochastic method was developed, with the incorporation of an evolutionary algorithm.

Deterministic approach

Aminoacid residue	Sequence position
Phenylalanine	42
Aspartate	23
Aspartate	71
Glutamate	234
Glycine	235
Alanine	277
Tyrosine	281
Isoleucine	282
Phenylalanine	55
Valine	83



>Wild-type
MPSIKLNSGYDMPAVGFGCWKVDVDTCSEQIYRAIKTGYRLFDGAEDYANEKLVGAGVKKAIDEGIVKREDLFLTLSDLQVD...
>Mutant-F42A
MPSIKLNSGYDMPAVGFGCWKVDVDTCSEQIYRAIKTGYRLADGAEDYANEKLVGAGVKKAIDEGIVKREDLFLTLSDLQVD...

Cofactor specificity prediction: Wild-type: NAD(H)
Mutant-F42A: NAD(H)

Cofactor specificity change **NOT** achieved

Mutant prediction score < threshold (0.8)

>Wild-type
MPSIKLNSGYDMPAVGFGCWKVDVDTCSEQIYRAIKTGYRLFDGAEDYANEKLVGAGVKKAIDEGIVKREDLFLTLSDLQVD...
>Mutant-F42A/D23S
MPSIKLNSGYDMPAVGFGCWKVSVDTCSEQIYRAIKTGYRLADGAEDYANEKLVGAGVKKAIDEGIVKREDLFLTLSDLQVD...

Cofactor specificity prediction: Wild-type: NAD(H)
Mutant-F42A/D23S: NAD(H)

Cofactor specificity change **NOT** achieved

Mutant prediction score < threshold (0.8)

>Wild-type
MPSIKLNSGYDMPAVGFGCWKVDVDTCSEQIYRAIKTGYRLFDGAEDYANEKLVGAGVKKAIDEGIVKREDLFLTLSDLQVD...
>Mutant-F42A/D23S/D71Y
MPSIKLNSGYDMPAVGFGCWKVSVDTCSEQIYRAIKTGYRLADGAEDYANEKLVGAGVKKAIDEGIVKREYLFLTLSDLQVD...

Cofactor specificity prediction: Wild-type: NAD(H)
Mutant-F42A/D23S/D71Y: NADP(H)

Cofactor specificity change **ACHIEVED**

Mutant prediction score ≥ threshold (0.8)

Figure 5.3 Exemplification of the deterministic method implementation for cofactor specificity conversion. Given the ordered list of most influential residues for cofactor specificity and by consulting the point-mutation selection chart, the method selects and accumulates different mutations until the reverted cofactor specificity prediction is achieved.

5.3.3 Cofactor specificity reversal – Stochastic method

With the selection of the ten most suitable mutable residue positions for cofactor specificity reversal, and given the possibility of each residue position being mutated by the remaining 19 aminoacids residues, it becomes clear the impossibility of predicting the cofactor specificity of every mutant combination. To overcome this issue, a stochastic method was developed through the implementation of an evolutionary algorithm. These optimization algorithms perform the evolution of a population by mimicking biologic events such as natural selection. Each individual in a population is evaluated through a fitness function and compared with newly generated individuals created by the application of reproduction operators to selected parents. As in nature, only the fittest individuals are allowed to continue in the population and reproduce [22].

Figure 5.4 is depicts a representation of the implemented evolutionary algorithms. Given the target aminoacid residue sequence and the list of 10 mutable positions, an initial population of 100 mutant aminoacid residue sequences (individuals) was generated, with random mutations in the available mutable positions. In this work, 5 evolutionary algorithms were implemented, with the only difference being the maximum candidate size allowed, varying between 1 and 5 mutations per individual. The optimization process was run for 100 generations. During each generation, the cofactor specificity of each individual was predicted using NiCofactor. After evaluating the entire population, the two best scoring individuals were maintained for the next generation, while the next best 50 undertook a recombination process, being 90% by crossover, where two individuals are crossed over to generate two new individuals, and 10% by mutation, where an individual's suitable aminoacid residue is randomly mutated. The remaining lowest scoring 48 individuals were discarded and newly generated individuals with random mutations in the available mutable positions were incorporated in the population. In the end of the optimization process, the five mutant sequences, containing 1 to 5 mutations, with the highest cofactor prediction score for the opposite cofactor were retrieved and outputted as result.

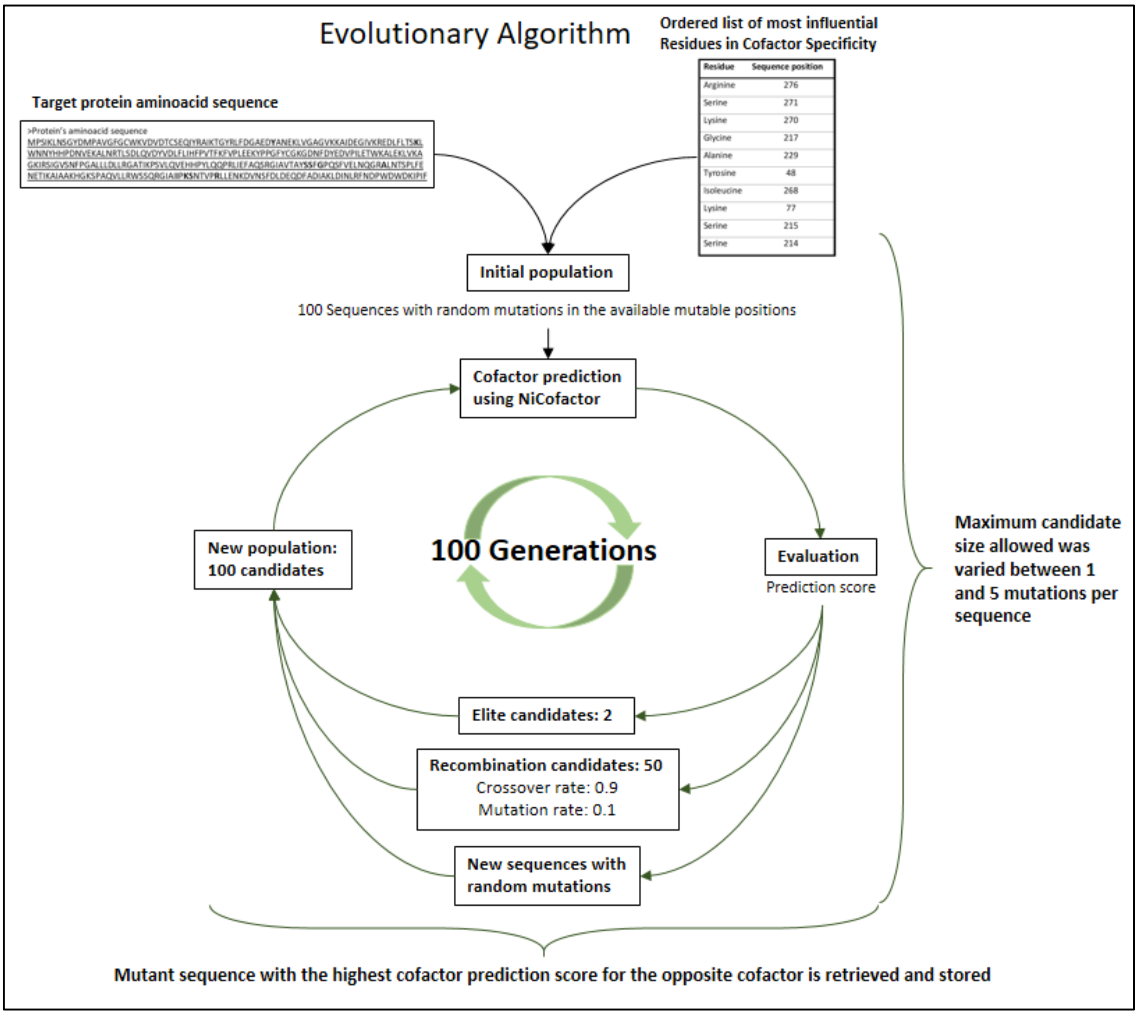

Evolutionary Algorithm
Target protein aminoacid sequence
>Protein's aminoacid sequence
MPSIKLNSGYDMPAVGFGCWKVDVDTCSEQIYRAIKTGYRLFDGAEDYANEKLVGAGVKKAIDEGIVKREDLFLTSKL
WNNYHHPDNVEKALNRTLSDLQVDYVDLFLIHFPVTFKFVPLEEKYPPGFYCGKGDNFDYEDVPILETWKALEKLVKA
GKIRSIGVSNFPGALLLDLLRGATIKPSVLQVEHHPYLQQPRLIEFAQSRGIAVTAYSSFGPQSFVELNQGRALNTSPLFE
NETIKAIAAKHGKSPAQVLLRWSSQRGIAIIPKSNTVPRLLENKDVNSFDLDEQDFADIAKLDINLRFNDPWDWDKIPIF
Ordered list of most influential Residues in Cofactor Specificity
Residue
Sequence position
Arginine
276
Serine
271
Lysine
270
Glycine
217
Alanine
229
Tyrosine
48
Isoleucine
268
Lysine
77
Serine
215
Serine
214
Initial population
100 Sequences with random mutations in the available mutable positions
Cofactor prediction using NiCofactor
100 Generations
New population: 100 candidates
Evaluation
Prediction score
Maximum candidate size allowed was varied between 1 and 5 mutations per sequence
Elite candidates: 2
Recombination candidates: 50
Crossover rate: 0.9
Mutation rate: 0.1
New sequences with random mutations
Mutant sequence with the highest cofactor prediction score for the opposite cofactor is retrieved and stored

Figure 5.4 – Evolutionary algorithm implemented in the development of the stochastic method for cofactor specificity reversal. During 100 generations, an optimization process is conducted, with multiple mutations being performed and analyzed with NiCofactor. Highest scoring mutants are retrieved and outputted as results.

5.3.4 Case studies

With the intent of assessing the performance of the developed methods for cofactor specificity reversal, four case studies were replicated *in silico* and their cofactor specificity reverted, using the developed methods. From the group of cofactor engineering studies published by Khoury and coworkers [23], analyzed in chapter 3, the four enzymes that were found to have completely reverted specificity or largely decreased affinity for one of the cofactors, increasing the affinity of the other, were selected as case studies. These were the cases of xylose reductase from *Picchia stipitis* (PsXR) [24], xylitol dehydrogenase from *Gluconobacter oxydans* (GoXD) [25], xylitol dehydrogenase from *Pichia stipitis* (PsXD) [26] and leucine dehydrogenase from *Tramitichromis intermedius* (TiLD) [27]. For these enzymes, NiCofactor was able to correctly predict the cofactor specificity of both wild-type and specificity reversed mutants. With that in mind, the four enzymes' aminoacid residue sequence were retrieved and processed using the above described methods with the intent of showcasing the results achieved *in silico*, and comparing them to the experimentally determined data on cofactor specificity reversing mutations.

From the four enzymes analyzed, three are part of the xylose metabolism, an extremely important pathway due to its great economical potential. Being a major component of hemicellulose and only second to glucose as the most abundant sugar in nature, D-xylose can be bioconverted from agricultural biomass wastes into biofuels, such as ethanol, through fermentation processes. However, *Saccharomyces cerevisiae*, the best adapted microorganism for producing ethanol, is not genetically equipped for metabolizing xylose. To solve this problem, xylose fermenting genes have been cloned in *S. cereviseae* from other organisms capable of metabolizing this sugar, such as *Pichia spitipis* [24], [26] and *Gluconobacter oxidans* [25]. Xylose reductase (EC 1.1.1.21) reduces xylose into xylitol using NADPH and xylitol dehydrogenase (EC 1.1.1.9) oxidizes, posteriorly, xylitol into xylulose, using NAD$^+$. Nonetheless, this difference in cofactor specificity creates an intercellular redox unbalance, hindering ethanol production yields and promoting xylitol excretion. An elegant solution implemented to solve this problem is the cofactor specificity reversal of xylose reductase from NADPH to NADH [24] or, by alternative, the specificity reversal of xylitol dehydrogenase from NAD$^+$ to NADP$^+$, taking advantage of the often higher availability of NADP$^+$ in the cell [25], [26].

The remaining enzyme, Leucine dehydrogenase from *Thermoactinomyces intermedius* uses NAD$^+$ for catalyzing the reversible deamination of L-leucine to its 2-oxo analogue, 4-methyl-2-oxopentanoate. As biosynthesis reactions generally use NADP$^+$ as cofactor, leucine dehydrogenase cofactor specificity reversal might improve this reaction's efficiency [27].

5.3.4.1 PsXR – *Pichia stipitis* xylose reductase

Xylose reductase (PsXR), from *Pichia stipitis*, was the only enzyme in the analyzed group with cofactor specificity for NADP(H), with the remaining enzymes being specific for NAD(H). In table 5.2 the results achieved for the *in silico* cofactor specificity change of PsXR are displayed.

As previously stated, NiCofactor was able to correctly predict the cofactor specificity from both wild-type and literature cofactor reversed mutant, being the results achieved by the deterministic and stochastic methods only outputted when the prediction score threshold is achieved.

Table 5.2. Mutations, cofactor predictions and prediction scores from literature experimental data, as well as from the implementation of both methods for reversing cofactor specificity *in silico* of PsXR. The deterministic method outputs only one mutant, while the stochastic method outputs five different mutants with the best found set of mutations for specificity reversal, according to the maximum candidate size allowed by the method, with the number on the gene name corresponding to the number of mutations selected.

Gene name	Mutation	Predicted cofactor	Prediction score
PsXR	Wild-type	NADP	0.8764
PsXR Literature	K270S/S271G/N272P/R276F	NAD	0.8792
PsXR Deterministic	S271E/R276E	NAD	0.9747
PsXR Stochastic 1	R276D	NAD	0.8661
PsXR Stochastic 2	K270D/S271D	NAD	0.9773
PsXR Stochastic 3	K270D/S271D/R276D	NAD	0.9996
PsXR Stochastic 4	S215R/K270D/S271D/R276D	NAD	0.9995
PsXR Stochastic 5	G217D/I268R / K270D/S271D/R276D	NAD	0.9999

When analyzing table 5.2 it is also possible to observe the amount and type of point mutations recommended by the developed methods in order to achieve cofactor specificity reversal. Being the literature mutant composed by four point-mutations, achieved through the implementation of a combinatorial active-site saturation mutagenesis method, we can observe that both deterministic and stochastic methods here implemented were able to predict mutants with fewer point-mutations and higher predicted reversed cofactor specificity. In this case, PsXR Deterministic is composed by only two point-mutations, with an Arginine (R) and a Serine (S) being substituted by a Glutamate (E). In the case of PsXR stochastic, the evolutionary algorithm was able to find a mutant with predicted reversed cofactor specificity with only one point-mutation, being this individual, due to its lower amount of point-mutations, considered the best hypothesis for performing *in vivo* cofactor specificity reversal. When analyzing the remaining stochastic mutants, it is observed that the mutants with higher amounts of point-mutations tend to incorporate the point-mutations predicted for the stochastic mutants with fewer point-mutations, indicating a strong effect of these mutations for the reversal of cofactor specificity. We can also see, in the stochastic mutants, that the cofactor prediction scores increase with the amount of point-mutations predicted, however, preference should be given to mutants with fewer mutations in order to preserve the structural stability of the enzyme.

Figure 5.5 depicts the positions, on the cofactor binding spot of PsXR, of the ten optimal mutable residues for cofactor specificity reversal in the wild-type, the point-mutations of the literature data, as well as the point-mutations predicted as the most suitable for achieving cofactor specificity reversal from both deterministic and stochastic methods.

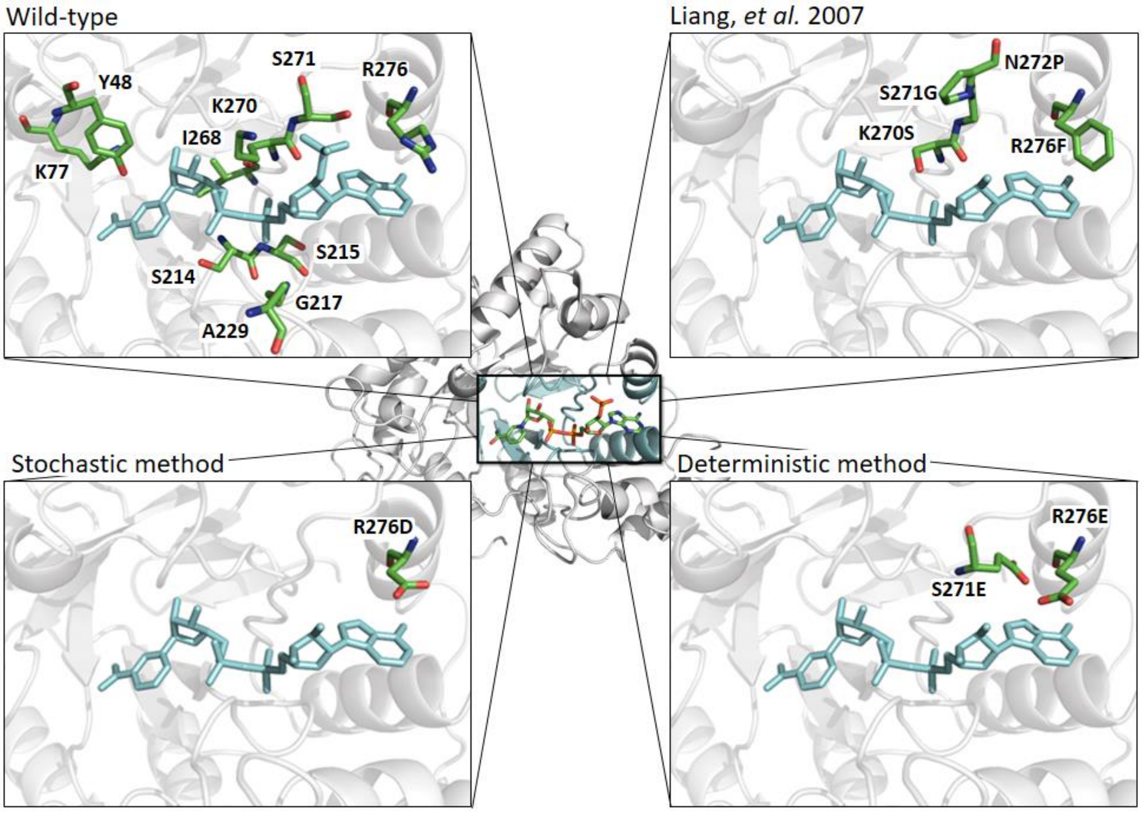


Figure 5.5 - Depiction of the structure of PsXR Wild-type, with emphasis on cofactor binding-pocket, showcasing the wild-type with optimal residues for specificity reversal (top-left), literature experimental data point-mutations position (top-right), best prediction from stochastic method (bottom-left) and prediction from deterministic method (bottom-right). NAD^{+} is represented in blue, whereas selected residues are green. Labels indicate the original residue/sequence position/mutation.

When analyzing figure 5.5 we can observe that residues selected as optimal for cofactor specificity reversal are dispersed in the cofactor binding-pocket, with greater incidence of Serine (S) and the negatively charged residues Arginine (R) and Lysine (K). By comparing the experimental data with the predictions performed, it is observed a close relation between *in vivo* and *in silico* results, with point-mutation predictions from both methods being in residues also targeted experimentally, as is the case of R276 and S271, both present near the adenine moiety of the cofactor. However, the type of residues selected as mutants *in silico* differ from the literature. The fact that the selected residues are the result of optimization methods using machine learning and evolutionary algorithms lead us to believe that these residues are better suited as specificity reversal point-mutations, which explains the lower amount of mutations required for specificity change.

5.3.4.2 GoXD –*Gluconobacter oxydans* xylitol dehydrogenase

Xylitol dehydrogenase (GoXD), from *Gluconobacter oxydans*, oxidizes xylitol into xylulose and has been shown to use exclusively NAD$^+$ as a reaction cofactor [25]. Table 5.3 displays the results achieved for the *in silico* cofactor specificity change of GoXD from NAD$^+$ dependent to NADP$^+$.

Table 5.3. Mutations, cofactor predictions and prediction scores from literature experimental data, as well as from the implementation of both methods for reversing cofactor specificity *in silico* of GoXD. The deterministic method outputs only one mutant, while the stochastic method outputs five different mutants with the best found set of mutations for specificity reversal, according to the maximum candidate size allowed by the method, with the number on the gene name corresponding to the number of mutations selected.

Gene name	Mutation	Predicted cofactor	Prediction score
GoXD	Wild-type	NAD	0.9678
GoXD Literature	D38S/M39R	NADP	0.9541
GoXD Deterministic	D38Y/A92Q/G93R	NADP	0.8822
GoXD Stochastic 1	D38R	NADP	0.8668
GoXD Stochastic 2	D38R/A92R	NADP	0.9752
GoXD Stochastic 3	D38R/D64R/A92R	NADP	0.9944
GoXD Stochastic 4	G16K/D38R/D64S/G93R	NADP	0.9957
GoXD Stochastic 5	G14S/G16T/D38R/D64R/A92R	NADP	0.9999

The results displayed in table 5.3 show a high score for the prediction of cofactor specificity of wild-type and literature experimental data, increasing the confidence level on the agreement between predicted and experimental results. In the analyzed case-study, the literature mutant was achieved with only two point-mutations, while the deterministic method required three mutations to achieve the same cofactor specificity prediction. As to GoXD stochastic results, the developed method was able to output a predicted reversed cofactor specificity mutant encompassing only one point-mutation, being it considered the best hypothesis for performing *in vivo* cofactor specificity reversal with minimal interventions. When further analyzing the achieved results, it is possible to observe that both the literature mutant and the selected stochastic mutant share similar features despite being originated from different approaches. From these approaches, the one described in the literature is the most laborious, involving structure characterization and structural alignment with other enzymes, together with the multiple selection of conserved aminoacid residue positions. In the literature mutant an Aspartate residue in position 38 was deleted and an Arginine residue was incorporated in position 39, whereas in the stochastic mutant, this event occurred in the same spot, position 38. A common characteristic that the predicted mutations appear to possess, in order to successfully reverting cofactor specificity in this case, is the promotion of Arginine (R) inclusion and the exclusion of Aspartate (D) in the analyzed sequences. This trait has been previously referred by Carugo and coworkers [19], in an early study on the subject, when observing conserved structural features of NADP-dependent enzymes, and was found throughout the analyzed case-studies in this work.

Figure 5.6 depicts the positions, on the cofactor binding spot of GoXD, of the ten optimal mutable residues for cofactor specificity reversal in the wild-type, the point-mutations of the literature data, as well as the point-mutations predicted as the most suitable for achieving cofactor specificity reversal from both deterministic and stochastic methods.

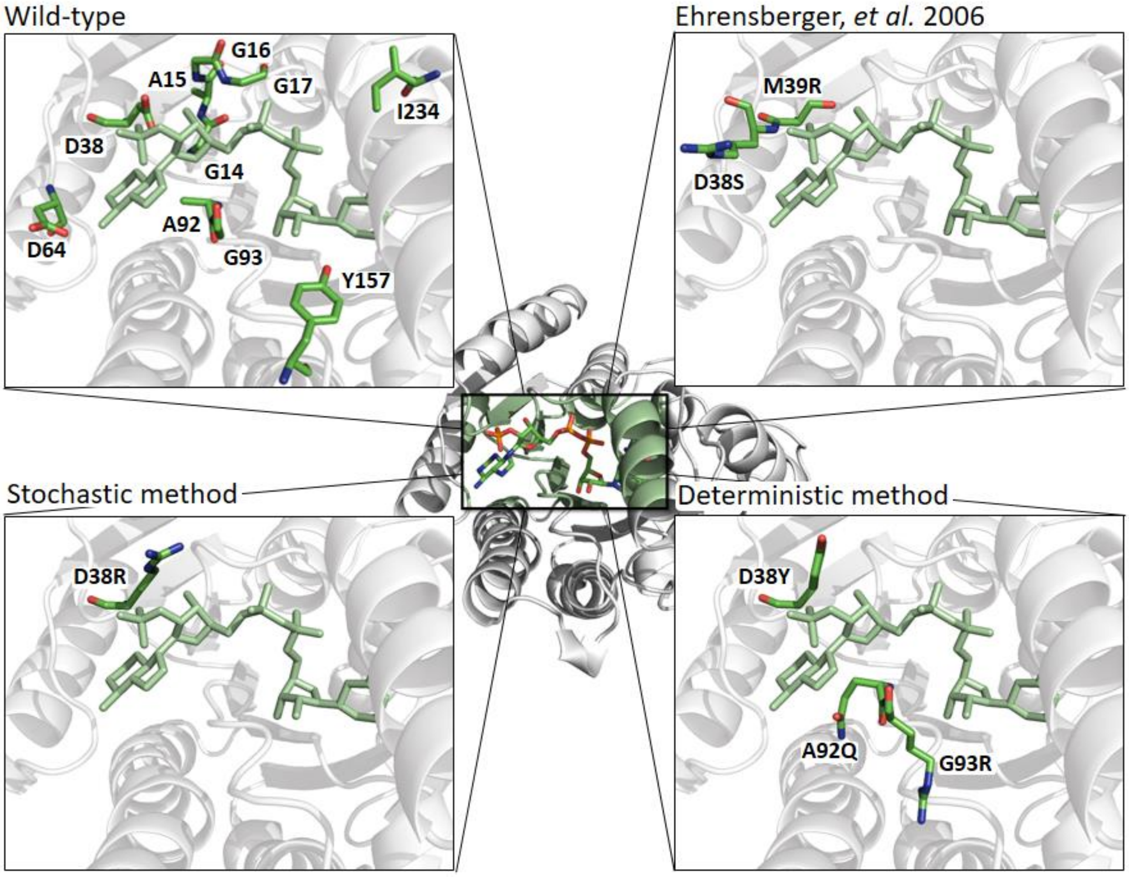


Figure 5.6 - Depiction of the structure of GoXD Wild-type, with emphasis on cofactor binding-pocket, showcasing the wild-type with optimal residues for specificity reversal (top-left), literature experimental data point-mutations position (top-right), best prediction from stochastic method (bottom-left) and prediction from deterministic method (bottom-right). NADP^{+} is represented in light green, whereas selected residues are green. Labels indicate the original residue/sequence position/mutation.

When analyzing figure 5.6 we can again observe that residues selected as optimal for cofactor specificity reversal are dispersed in the cofactor binding-pocket; however, this time with greater incidence of Aspartate (D) near the adenine moiety, and Glycine (G). Point-mutations from both literature and selected stochastic mutants are located near the 2'-phosphate, while in the deterministic mutant, point-mutations are also present below the cofactor structure, near the adenine and the ribose in the adenine moiety.

5.3.4.3 PsXD –*Pichia stipitis* xylitol dehydrogenase

Xylitol dehydrogenase (PsXD), from *Pichia stipitis*, was the second xylitol dehydrogenase analyzed in this study. In table 5.4 are displayed the literature and the *in silico* results achieved for the cofactor specificity reversal of PsXD.

Table 5.4. Mutations, cofactor predictions and prediction scores from literature experimental data, as well as from the implementation of both methods for reversing cofactor specificity *in silico* of PsXD. The deterministic method outputs only one mutant, while the stochastic method outputs five different mutants with the best found set of mutations for specificity reversal, according to the maximum candidate size allowed by the method, with the number on the gene name corresponding to the number of mutations selected.

Gene name	Mutation	Predicted cofactor	Prediction score
PsXD	Wild-type	NAD	0.9816
PsXD Literature	D207A/I208R/F209S/N211R	NADP	0.9906
PsXD Deterministic	G183R/G185R/V187A/D207Y/A254Q/V274A	NADP	0.8806
PsXD Stochastic 1	D207R	NADP	0.8198
PsXD Stochastic 2	D207Y/I208R	NADP	0.9890
PsXD Stochastic 3	D207S/I208R/F209K	NADP	0.9944
PsXD Stochastic 4	D207S/I208R/F209Y/A254R	NADP	0.9999
PsXD Stochastic 5	G183K/D207R/I208R/A254R/V274Y	NADP	0.9999

Once again, when analyzing the prediction scores displayed in table 5.4, both wild-type and the literature mutant show a high prediction score. This time, however, four point-mutations were required in order to perform the cofactor specificity change in the literature mutant, achieved by site directed mutagenesis. By comparing both xylitol dehydrogenases analyzed, we can observe that *Pichia stipitis*' xylitol dehydrogenase requires twice as many mutations to perform the same specificity change as *Gluconobacter oxydans*' xylitol dehydrogenase. Despite being able to catalyze the same reaction, these enzymes are different, being originated from two different organisms, which explains the different approaches needed to perform the same conversion. In this case study, the deterministic method required the mutation of six residues before being able to effectively predict cofactor specificity change in PsXD, also demonstrating the challenging task of performing cofactor specificity reversing PsXD. The stochastic method, on the other hand, and despite the large amount of point-mutations required in the previous two PsXD mutant results, was again able to predict cofactor specificity reversal with the implementation of only one point-mutation. However, when comparing its prediction score to the remaining stochastic mutants, and despite being above the prediction threshold, it is observed a significant increase in prediction scores in mutants with more point-mutations.

Figure 5.7 depicts the positions of the optimal mutable residues for cofactor specificity reversal, as well as the point-mutations from the literature data and both deterministic and stochastic methods, on the cofactor binding spot of PsXD.

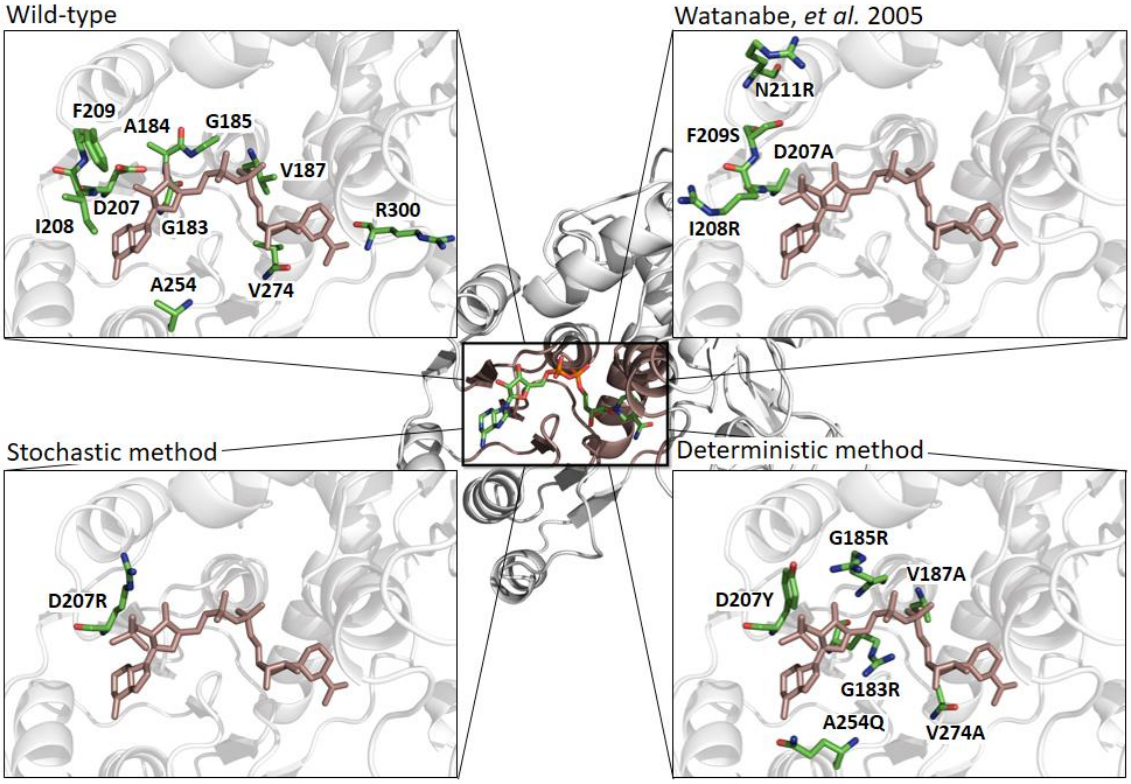


Figure 5.7 - Depiction of the structure of PsXD Wild-type, with emphasis on cofactor binding-pocket, showcasing the wild-type with optimal residues for specificity reversal (top-left), literature experimental data point-mutations position (top-right), best prediction from stochastic method (bottom-left) and prediction from deterministic method (bottom-right). NADP$^+$ is represented in light red, whereas selected residues are green. Labels indicate the original residue/sequence position/mutation.

When comparing the results depicted in the figure, it is again possible to observe some similarities between the literature and the stochastic method. Despite the literature mutant requires more point-mutations in order to achieve cofactor specificity reversal, the underlying mechanism is in some part shared with the results achieved by the stochastic method, with negatively charged residues being replaced by positive ones in the 2'-phosphate area, among other mutations. This serves to showcase the optimization and precision achieved by the developed methods. Instead of experimenting with a large amount of random mutations trying to almost blindly achieve cofactor specificity reversal as with some experimental methods, such us random or combinatorial mutagenesis, the developed methods optimize the search space in order to encounter the optimal set of mutations required to perform cofactor specificity.

5.3.4.4 TiLD – *Thermoactinomyces intermedius* leucine dehydrogenase

Leucine dehydrogenase, from *Thermoactinomyces intermedius*, the only enzyme in this case study not involved in xylose metabolism, depends on NAD(H) to catalyze its reaction. Table 5.5 displays the results achieved for the *in silico* cofactor specificity reversal of TiLD, as well as the results of wild-type and literature mutant cofactor specificity prediction.

Table 5.5. Mutations, cofactor predictions and prediction scores from literature experimental data, as well as from the implementation of both methods for reversing cofactor specificity *in silico* of TiLD. The deterministic method outputs only one mutant, while the stochastic method outputs five different mutants with the best found set of mutations for specificity reversal, according to the maximum candidate size allowed by the method, with the number on the gene name corresponding to the number of mutations selected.

Gene name	Mutation	Predicted cofactor	Prediction score
TiLD	Wild-type	NAD	0.9776
TiLD Literature	D203A/I204R/D210R	NADP	0.9546
TiLD Deterministic	D203Y/V116A/A238Q/G180R	NADP	0.8607
TiLD Stochastic 1	D203R	NADP	0.8534
TiLD Stochastic 2	D203R/I204R	NADP	0.9954
TiLD Stochastic 3	A238R/D203S/I204R	NADP	0.9969
TiLD Stochastic 4	I204R/D203S/G180R/A238Y	NADP	0.9999
TiLD Stochastic 5	A185T/E114Q/D203R/I204R/A238R	NADP	0.9999

The results displayed in table 5.5 for the prediction of experimental data on wild-type and mutant cofactor specificity show, once again, a very high prediction score. In fact, it is possible to observe that, from the four enzymes analyzed in this case study, all three enzymes originally binding to NAD(H) have a higher prediction score than *Pichia stipitis* xylose reductase, the only enzyme in the study originally binding to NADP(H). In this case, three point-mutations were required to achieve cofactor specificity reversal in the literature mutant using the systematic replacement of target aminoacid residues, with the deterministic approach's mutant requiring four, according with the developed method. Concerning the stochastic approach, once again a mutant encompassing only one mutation was able to be predicted, with this approach having successfully identified single mutants with predicted reverse cofactor specificities for all cases presented. The behavior of the remaining stochastic mutants is similar to the previously analyzed case-studies, where mutants with a higher number of mutations have a higher cofactor specificity prediction score.

Figure 5.8 depicts the optimal mutable residues for cofactor specificity reversal, as well as the point-mutations from the literature data and both deterministic and stochastic methods, on the cofactor binding spot of TiLD.

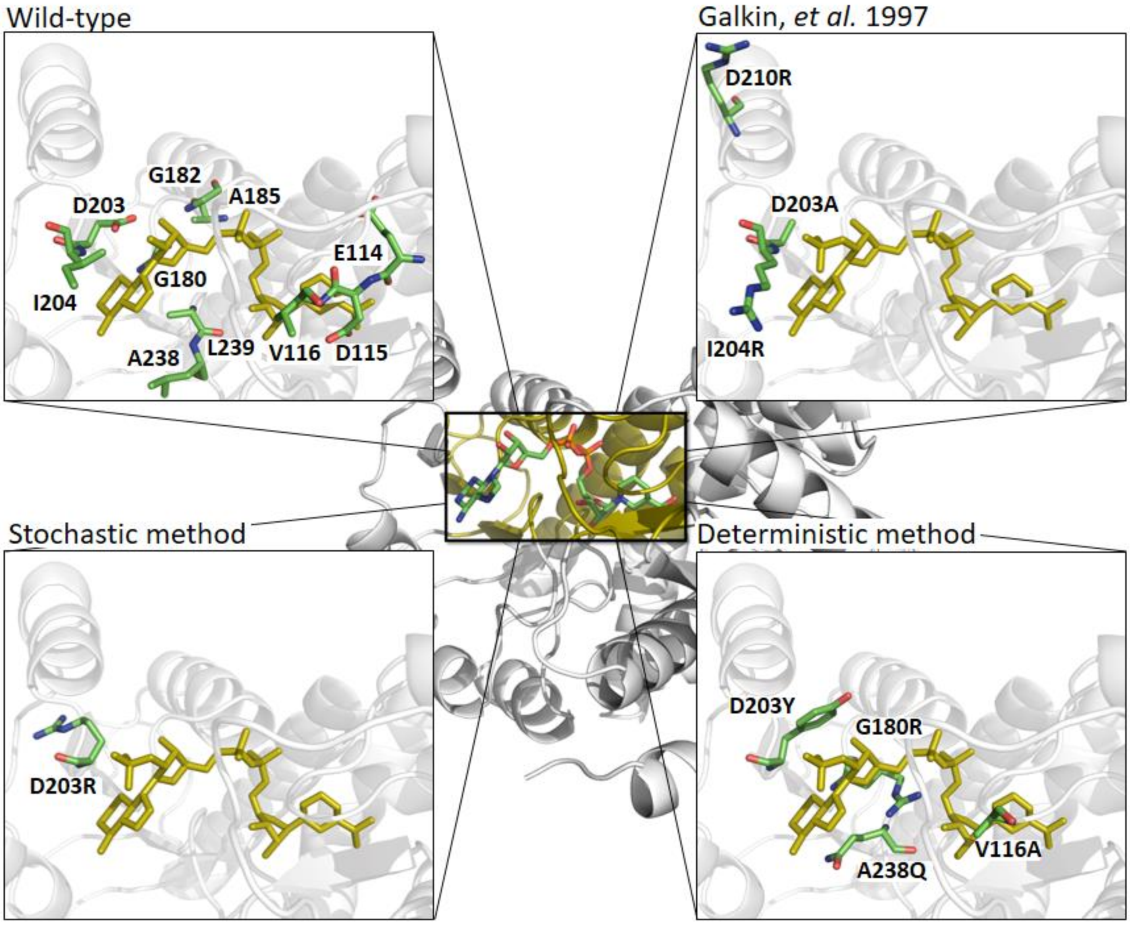


Figure 5.8 - Depiction of the structure of TiLD Wild-type, with emphasis on cofactor binding-pocket, showcasing the wild-type with optimal residues for specificity reversal (top-left), literature experimental data point-mutations position (top-right), best prediction from stochastic method (bottom-left) and prediction from deterministic method (bottom-right). NADP^{+} is represented in yellow, whereas selected residues are green. Labels indicate the original residue/sequence position/mutation.

As it is possible to observe in figure 5.8, the residues identified as optimal for cofactor specificity change mutations, are dispersed in the cofactor binding site. From the three mutant residues undergone point-mutations reported in the literature, two were also predicted as capable of reverting cofactor specificity by the developed methods, Aspartate in position 203 (D203) and Isoleucine in position 204 (I204). However, in the resulting mutants outputted by the developed methods, only D203 was used for the selected best mutant, being its mutation by an Arginine residue (D203R), in the stochastic method, sufficient for predicting an altered cofactor specificity. In the mutant outputted by the deterministic method, besides the mutation of D203 by a tyrosine residue (D203Y), three other mutations were necessary in order to predict a mutant with reversed cofactor specificity (V116A/G180R/A238Q).

5.4 Conclusions

The molecular determinants responsible for NAD(P)(H) cofactor specificity are still regarded as illusive and difficult to characterize. With metabolic engineering and strain design endeavors increasingly requiring more rational approaches to maximize efficiency and reduce costs, new methods for fast and accurate large scale data processing are in need. In a previous chapter, having moved a step forward in unveiling the molecular determinants for cofactor specificity, using enzyme structural information, we presented in this chapter *in silico* methods for the conversion of NAD(P)(H) cofactor specificity in structurally uncharacterized enzymes, with the identification of optimal residues for mutation and the set of mutations better suited for performing such conversion. Learning from the data generated in previous chapters, we were able to elaborate and test two methods for the automation of cofactor engineering reversal through computational experiments, drastically reducing the large combinatorial space of mutations available. Using the previously developed prediction tool NiCofactor, such mutation predictions were promptly tested and analyzed. In the first approach, by switching residues with a large impact on specificity, in a rational way, with residues with large impact on specificity for the opposite cofactor, we found a viable approach to suggest cofactor specificity changing mutations capable of disrupting the original cofactor specificity and revert it. On a second method, a stochastic approach was implemented in order to maximize the representation of the combinatorial space generated by the identification of optimal mutations for specificity reversal. Using evolutionary algorithms, several mutants were constructed and predicted, with the fittest feeding new generations of mutants in the optimization process.

The results presented in this chapter were complemented by the analysis of four different case studies reported in previous chapters and successfully replicated *in silico*, through the prediction of cofactor specificity using NiCofactor. For each of the analyzed case-studies, both developed methods were able to predict optimal mutants with reversed cofactor specificities. Several of the mutant sequences outputted by the methods have fewer point-mutations than the ones observed in the experimental data. From the two developed methods, the stochastic approach was able to produce the best results for achieving cofactor specificity reversal, with the output of 5 different solutions for every case-study and always presenting the least amount of predicted point-mutations necessary to perform specificity change. In fact, for each of the four case-studies, the stochastic method was able to predict and output a mutant sequence, with reversed cofactor specificity prediction, containing only one point-mutation. In the case of the deterministic approach, this method was also able to predict, for each case-study, a mutant sequence with reversed cofactor specificity, but with a higher amount of point-mutations, varying between two and six in order to revert specificity.

As for the location of the cofactor specificity reverting point-mutations, both methods, as well as the experimental data, appear to prioritize the adenine moiety of the cofactor as the receiving area of mutations, in the binding pocket of both cofactors. In all cases, it is also recurrent the occurrence of mutations involving the alteration of positively charged residues (Arginine or Lysine) for negative (Aspartate or Glutamate) and vice-versa, as first reported by Carugo, *et al.* in 1997 [19], for conserved structures, in order to reverse cofactor specificity.

The developed methods showed a good coverage of the addressed problem and combinatorial solution space available, being able to perform predictions and find optimal subjects for reversing cofactor specificity, optimizing, *in silico*, the arduous and expensive task of performing random mutagenesis, the most common approach taken on this problem. However, it is important to note that, despite the good confidence levels reported during the development of this work, the cofactor specificity reversal claims made here refer to *in silico* predictions and are subjected to the accuracy levels achieved. The fact that a 90% accuracy is claimed for the correct prediction of cofactor specificity means that there is a margin for inaccurate outputted results. Also, structural inference methods using comparative modeling are also not 100% representative of the natural enzyme's structure, despite their usefulness and accuracy. These facts, together with inherent structural destabilization that a mutation might induce, lead us to believe that a thorough downstream enzymatic activity assession is required in order to validate the developed methods and obtain *in vivo* successful results.

Nonetheless, the presented *in silico* results evidence a breakthrough, not only for NAD(P)(H) cofactor specificity problems, but also in the approach to be taken in new problems involving different types of substrates or cofactors, taking great advantage of the structural information stored in an enzyme 3D structure. We believe these results are of great utility and represent an important contribution for cofactor engineering problems, rational metabolic engineering and strain design endeavors.

5.5 References

[1] B. M. Woolston, S. Edgar, and G. Stephanopoulos, "Metabolic Engineering: Past and Future," *Annu. Rev. Chem. Biomol. Eng.*, vol. 4, no. 1, pp. 259–288, 2013.

[2] U. T. Bornscheuer, G. W. Huisman, R. J. Kazlauskas, S. Lutz, J. C. Moore, and K. Robins, "Engineering the third wave of biocatalysis," *Nature*, vol. 485, no. 7397, pp. 185–194, 2012.

[3] B. Ø. Palsson, "Metabolic Systems Biology," *FEBS Lett.*, vol. 583, no. 24, pp. 3900–3904, 2011.

[4] J. E. Bailey, "Toward a Science of Metabolic Engineering," *Science (80-.).*, vol. 252, pp. 1668–1697, 1991.

[5] C. Willrodt, R. Karande, A. Schmid, and M. K. Julsing, "Guiding efficient microbial synthesis of non-natural chemicals by physicochemical properties of reactants," *Curr. Opin. Biotechnol.*, vol. 35, pp. 52–62, 2015.

[6] W. Liu and P. Wang, "Cofactor regeneration for sustainable enzymatic biosynthesis," *Biotechnol. Adv.*, vol. 25, no. 4, pp. 369–384, 2007.

[7] Y. Wang, K. Y. San, and G. N. Bennett, "Cofactor engineering for advancing chemical biotechnology," *Curr. Opin. Biotechnol.*, vol. 24, no. 6, pp. 994–999, 2013.

[8] J. Park and Y. Choi, "Cofactor engineering in cyanobacteria to overcome imbalance between NADPH and NADH: A mini review," *Front. Chem. Sci. Eng.*, vol. 11, no. 1, pp. 66–71, 2017.

[9] L. S. Vidal, C. L. Kelly, P. M. Mordaka, and J. T. Heap, "Review of NAD (P) H-dependent oxidoreductases : Properties , engineering and application," *BBA - Proteins Proteomics*, vol. 1866, no. August 2017, pp. 327–347, 2018.

[10] R. R. Bommareddy, Z. Chen, S. Rappert, and A. P. Zeng, "A de novo NADPH generation pathway for improving lysine production of Corynebacterium glutamicum by rational design of the coenzyme specificity of glyceraldehyde 3-phosphate dehydrogenase," *Metab. Eng.*, vol. 25, pp. 30–37, 2014.

[11] Z. Xiao, C. Lv, C. Gao, J. Qin, C. Ma, Z. Liu, P. Liu, L. Li, and P. Xu, "A novel whole-cell biocatalyst with NAD$^+$ regeneration for production of chiral chemicals," *PLoS One*, vol. 5, no.

1, pp. 1–6, 2010.

[12] H. TAMAKAWA, S. IKUSHIMA, and S. YOSHIDA, “Ethanol Production from Xylose by a Recombinant *Candida utilis* Strain Expressing Protein-Engineered Xylose Reductase and Xylitol Dehydrogenase,” *Biosci. Biotechnol. Biochem.*, vol. 75, no. 10, pp. 1994–2000, 2011.

[13] S. Bastian, X. Liu, J. T. Meyerowitz, C. D. Snow, M. M. Y. Chen, and F. H. Arnold, “Engineered ketol-acid reductoisomerase and alcohol dehydrogenase enable anaerobic 2-methylpropan-1-ol production at theoretical yield in Escherichia coli,” *Metab. Eng.*, vol. 13, no. 3, pp. 345–352, 2011.

[14] C. R. Shen, E. I. Lan, Y. Dekishima, A. Baez, K. M. Cho, and J. C. Liao, “Driving forces enable high-titer anaerobic 1-butanol synthesis in Escherichia coli,” *Appl. Environ. Microbiol.*, vol. 77, no. 9, pp. 2905–2915, 2011.

[15] R. Pereira, J. Nielsen, and I. Rocha, “Improving the flux distributions simulated with genome-scale metabolic models of Saccharomyces cerevisiae,” *Metab. Eng. Commun.*, vol. 3, pp. 153–163, 2016.

[16] Z. A. King and A. M. Feist, “Optimizing Cofactor Specificity of Oxidoreductase Enzymes for the Generation of Microbial Production Strains—OptSwap,” *Ind. Biotechnol.*, vol. 9, no. 4, pp. 236–246, 2013.

[17] D. Cui, L. Zhang, S. Jiang, Z. Yao, B. Gao, J. Lin, Y. A. Yuan, and D. Wei, “A computational strategy for altering an enzyme in its cofactor preference to NAD(H) and/or NADP(H),” *FEBS J.*, vol. 282, no. 12, pp. 2339–2351, 2015.

[18] J. K. B. Cahn, C. A. Werlang, A. Baumschlager, S. Brinkmann-Chen, S. L. Mayo, and F. H. Arnold, “A General Tool for Engineering the NAD/NADP Cofactor Preference of Oxidoreductases,” *ACS Synth. Biol.*, vol. 6, no. 2, pp. 326–333, 2017.

[19] O. Carugo and P. Argos, “NADP-dependent enzymes. I: Conserved stereochemistry of cofactor binding,” *Proteins Struct. Funct. Genet.*, vol. 28, no. 1, pp. 10–20, 1997.

[20] J. Barhak and A. Garrett, “Population Generation from Statistics Using Genetic Algorithms with MIST + INSPYRED,” *MODSIM World 2014*, pp. 1–8, 2014.

[21] W. DeLano, “Pymol: An open-source molecular graphics tool,” *CCP4 Newsl. Protein Crystallogr.*, vol. 700, 2002.

[22] Z. Michalwicz, “Evolutionary Programming and Genetic Programming,” *Genet. Algorithms + Data Struct. = Evol. Programs*, vol. 13, pp. 283–287, 1996.

[23] G. A. Khoury, H. Fazelinia, J. W. Chin, R. J. Pantazes, P. C. Cirino, and C. D. Maranas, “Computational design of Candida boidinii xylose reductase for altered cofactor specificity,” *Protein Sci.*, vol. 18, no. 10, pp. 2125–2138, 2009.

[24] L. Liang, J. Zhang, and Z. Lin, “Altering coenzyme specificity of Pichia stipitis xylose reductase by the semi-rational approach CASTing,” *Microb. Cell Fact.*, vol. 6, pp. 1–11, 2007.

[25] A. H. Ehrensberger, R. A. Elling, and D. K. Wilson, “Structure-guided engineering of xylitol dehydrogenase cosubstrate specificity,” *Structure*, vol. 14, no. 3, pp. 567–575, 2006.

[26] S. Watanabe, T. Kodaki, and K. Makino, “Complete reversal of coenzyme specificity of xylitol dehydrogenase and increase of thermostability by the introduction of structural zinc,” *J. Biol. Chem.*, vol. 280, no. 11, pp. 10340–10349, 2005.

[27] a Galkin, L. Kulakova, T. Ohshima, N. Esaki, and K. Soda, “Construction of a new leucine dehydrogenase with preferred specificity for $NADP^+$ by site-directed mutagenesis of the strictly NAD^+-specific enzyme.,” *Protein Eng.*, vol. 10, no. 6, pp. 687–690, 1997.

CHAPTER 6

General conclusions and future work

The general purpose of the conducted research present in this thesis was the improvement of predictions in metabolic engineering problems by incorporating enzyme structural information. More specifically, this thesis focus in the problem of accurately identifying enzyme's usage and specificity for the cofactors nicotinamide adenine dinucleotide (NAD) and nicotinamide adenine dinucleotide phosphate (NADP). To achieve the proposed intention, specific research aims were elaborated in chapter 1. The main conclusions achieved with the pursuit of the specified research aims are presented below.

In chapter 2, an extensive review of the literature, regarding the state-of-the-art on methodologies for performing genome and enzyme functional annotations, was performed. Multiple concepts across different fields of the biological sciences were described and analyzed, from genome sequencing to metabolic engineering and strain design approaches using systems biology. It became evident that further development of novel structure-based methods was in need to understand and overcome several hurdles still encountered in the reviewed processes.

Being the molecular characterization of NAD(P)(H) cofactor specificity regarded as an extremely challenging task and due to its importance in metabolic and protein engineering problems, in chapter 3, the molecular determinants for NAD(P)(H) cofactor specificity were unveiled. Using enzyme structural information and machine learning algorithms we were able to identify responsible molecular determinants for cofactor specificity and apply the proposed findings to enzymes not structurally characterized. Support vector machines were used to classify a formulated problem using information on cofactor neighbor aminoacid residues. A group of 921 protein structures correspondent to enzymes bound to NAD(H) or NADP(H) were correctly identified with an accuracy of 96.20% using the stated method. A prediction tool was further developed and allowed the correct identification of cofactor specificity in a curated dataset of structurally uncharacterized enzymes with an accuracy of 83.5%. When applied a prediction score threshold of 0.8, 73.4% of the predictions were included, with a prediction accuracy of 90%. When the score threshold was increased to 0.95, nearly 50% of the predictions were present, with a prediction accuracy of 96%. A webserver was developed to allow a fast and user-friendly access to the automatic prediction of NAD(P)(H) cofactor specificity enzymes without structural characterization.

In chapter 4 the curation of genome-scale metabolic models (GEM) through the characterization of reactions using the cofactors NAD(P)(H) was performed. Being GEM reconstruction curation process a laborious and skill requiring task, reaction cofactor usage prediction in GEM reconstruction has an immediate effect on reaction composition. Using the developed cofactor prediction tool, 59 different reconstructed GEMs belonging to 47 different strains were analyzed using the aminoacid sequence of enzymes associated to reactions using NAD(P)(H). Results help depicting the state of cofactor curation in GEM reconstruction, with an overall satisfactory curation of NAD(P)(H) usage but encompassing cofactor usage mismatches that can impair an accurate GEM simulation. NAD(P)(H) using reactions were corrected in the most recent *Saccharomyces cerevisiae* GEM, Yeast 7.6, and the model was used in simulations. The results achieved show that the corrections implemented, by cofactor curation, improve the simulation performance of the model, being in a higher agreement with experimental data found in literature.

With the developments achieved for the prediction of cofactor specificity, in chapter 5 a new method for the *in silico* conversion of NAD(P)(H) cofactor specificity in enzymes not structurally characterized is proposed. With the data generated in previous chapters we elaborated and tested a hypothesis for the automation of cofactor specificity reversal, drastically reducing the large combinatorial space of mutations available. This was achieved by predicting the effect of switching aminoacid residues with large specificity impact. For that, two distinct approaches were implemented, being one deterministic, using the gathered information on the most influential residues for both cofactor specificities; and the other stochastic, using evolutionary algorithms to locate the optimal set of mutations capable of reverting cofactor specificity. Given the satisfactory results, new optimization steps are suggested to be implemented in the future in order to more accurately perform the proposed task. With metabolic engineering and strain design endeavors increasingly requiring more rational approaches and process automation, the presented developments are believed to be of great utility and represent an important contribution for cofactor engineering problems and metabolic engineering overall.

Future perspectives of the conducted research focus on the continuous improvement of the developed methodologies presented, as well as the broader application of said methodologies. Concerning the cofactor specificity prediction tool, future work includes the continuous development of the created tool through the integration of additional curated data on enzyme structures encompassing information on cofactor specificity. The integrated data shall be used to improve the existing prediction models, as well as in the optimization of the comparative modelling methods used to analyzed enzymes without structural characterization. Different machine learning methods should also be tested and implemented and their performance assessed.

The general approach developed in this thesis to solve the NAD(P)(H) specificity problem using structural information and machine learning algorithms should also be applied in solving different metabolic engineering and systems biology problems concerning the specificity of other cofactors, as well as substrate affinity and inhibitor recognition.

Regarding the application of the developed prediction tool in the curation of GEMs NAD(P)(H) using reactions, reaction correction and model simulations should also be performed in different GEMs and the resulting performance analyzed to assess their enhancement. Gene knock out and overexpression for the production of compounds of interest should be further simulated in the curated models and compared with experimental data also to evaluate model improvement.

As to the proposed method for the *in silico* conversion of NAD(P)(H) cofactor specificity in enzymes not structurally characterized, the predicting capabilities of the method should be further developed with the integration of new approaches for assessing enzyme's viability with the proposed mutations. The predicted cofactor reversed enzyme mutants resulting from the analyzed case studies in this thesis should be experimentally characterized and evaluated for their cofactor specificity. New candidate enzymes should be sought and their cofactor specificity conversion predicted.

APPENDIX

Table A1 – The presented table encompasses all extracted feature weights from the developed SVM predictive model in chapter 3.

NADP	weight	NADP	weight	NADP	weight	NADP	weight
O2B_ARG	0.449913142	O3B_MET	0.079733928	O1N_LYS	0.048374441	PN_PHE	0.020487316
O2B_SER	0.265379606	C5A_LEU	0.079072527	C5D_ARG	0.048085877	PN_VAL	0.020312879
O2B_LYS	0.261630434	O5D_HIS	0.078422943	O2N_PRO	0.048039874	C6A_ARG	0.020247682
C2B_ARG	0.228622965	N3A_GLY	0.078126706	C1B_GLY	0.047830205	C5A_ASN	0.020164465
O3_GLY	0.222471342	C2D_ILE	0.077926803	O4D_PHE	0.047278498	C2A_PHE	0.019981824
O3B_ALA	0.217947335	O1A_ILE	0.07767809	O5B_GLY	0.047202217	C3D_THR	0.019211148
C4D_ASN	0.21647244	C3N_GLY	0.077661844	C2N_TRP	0.046976252	O3_PRO	0.019170498
O1A_ALA	0.186032894	PA_VAL	0.076984447	O3_TRP	0.046775668	C1D_TYR	0.01907725
O2B_GLY	0.180156852	O4B_TYR	0.076807191	C4A_GLN	0.046168774	O4D_CYS	0.019057066
N1A_SER	0.167953722	C8A_ALA	0.076797825	O2B_TYR	0.046021503	PN_GLN	0.018273416
C5B_LEU	0.164328457	C1D_ASN	0.076774549	O5B_ILE	0.045863598	O2A_TYR	0.01808518
C1B_ARG	0.163757278	C3N_SER	0.076468785	C1B_TYR	0.045550284	N1N_ARG	0.017754525
O3D_TYR	0.161156286	O3D_ASP	0.076355444	N1A_ALA	0.045532792	N3A_HIS	0.017404048
O7N_CYS	0.15867393	O1N_TRP	0.07594696	O2A_PHE	0.04545846	O5B_MET	0.016805257
PN_GLU	0.158627537	O3B_ASN	0.07554993	C3D_MET	0.045343175	O4D_MET	0.016554396
C2N_ILE	0.157813268	N3A_ASN	0.07451086	C2A_VAL	0.044915605	C2D_TYR	0.016524546
C5A_TYR	0.157211897	N7N_LYS	0.074129083	C4B_THR	0.043327963	O2D_GLY	0.016109269
O2N_ASP	0.152875829	C4D_LYS	0.07406761	C5N_HIS	0.043243281	O3_TYR	0.015763729
C3N_LYS	0.149656203	C4B_TYR	0.073339554	N7A_ALA	0.04306512	N3A_VAL	0.01547717
O5D_ILE	0.14931893	N6A_ILE	0.073237285	C1D_LYS	0.042897475	O4D_HIS	0.015401676
C4D_THR	0.148308899	C5D_PRO	0.073147588	O5D_TRP	0.042875009	C3B_PHE	0.015107134
N7A_TYR	0.148010431	C1B_THR	0.072608212	O3D_LEU	0.042203353	C4B_SER	0.015050321
C2B_LYS	0.146967824	O2D_PRO	0.07220358	N9A_GLY	0.041931586	O5D_TYR	0.015017127
O1A_SER	0.145495323	C2N_VAL	0.0718949	N1A_ILE	0.04163029	O1A_GLU	0.01475175
C6A_VAL	0.145139302	C6N_ASN	0.071586964	PA_ASN	0.041202421	C8A_GLY	0.014641457
C4N_LEU	0.144604686	O7N_MET	0.071435417	C5A_SER	0.041119304	C3B_LYS	0.014594822
C4A_ASN	0.1366117	C5D_ASN	0.070929606	O7N_SER	0.040917906	C4D_MET	0.014577395
N9A_TYR	0.136561537	C6A_GLY	0.070840069	C2A_MET	0.040509358	PN_PRO	0.014217951
O2D_ALA	0.135126711	C4D_ILE	0.070215604	C6A_TYR	0.040119304	C2A_CYS	0.01413095
O2A_ALA	0.134424506	O3_ALA	0.069690483	C5A_GLN	0.039823784	C5D_TRP	0.014000081
O5B_SER	0.13329123	O1N_HIS	0.069401062	PA_ALA	0.039247128	O2A_VAL	0.013853835
N1A_PRO	0.131819349	O3D_TRP	0.068758458	C6N_ARG	0.03923571	O2A_THR	0.013831904
N7N_ASN	0.128466426	N9A_THR	0.068569727	C5N_MET	0.039171159	O3D_ARG	0.013697525
O1A_LYS	0.127108198	C6N_ALA	0.068517968	N7A_ASN	0.039122243	O2D_PHE	0.013391022
N7N_ALA	0.126482461	C7N_GLN	0.068492631	C5B_ILE	0.039102077	C5B_TYR	0.013353805
O3B_LYS	0.123755873	PA_TYR	0.067210455	C6A_SER	0.038986334	C1D_TRP	0.013208285
C5D_ALA	0.123037613	O2A_LEU	0.066556605	O2B_GLN	0.038486753	O5B_PHE	0.012902275
C2B_SER	0.122966829	O4D_TYR	0.06595204	C5B_GLY	0.038076815	O3B_PRO	0.012638936
C4N_GLY	0.11898915	O4B_ARG	0.065935256	O2B_ILE	0.037717059	N7N_SER	0.012619455
C8A_TYR	0.116865395	C4N_ARG	0.065840612	O1A_CYS	0.037609135	C4N_ASP	0.012426783
O3_ASN	0.115743489	C4N_SER	0.065697869	O3D_LYS	0.036670257	C4N_MET	0.012024085
O4D_THR	0.11469516	O2N_GLN	0.065514055	C5D_TYR	0.036588347	O2B_TRP	0.011954468
C4A_ALA	0.114322444	C1D_GLU	0.065489345	C2N_LYS	0.036131842	C2D_SER	0.011876911
C4B_LEU	0.113631168	C4N_HIS	0.065426674	C2B_PHE	0.035967943	N3A_PRO	0.011724681
N1N_ASP	0.113602654	O3D_HIS	0.064528665	C5D_ILE	0.035936608	C1D_ARG	0.011535813
C5B_CYS	0.113187311	C5N_SER	0.064480807	O2D_ASP	0.035202961	C5B_PHE	0.01140188
O4B_GLY	0.111937545	C3D_PRO	0.064476088	O4B_VAL	0.035068685	O2A_HIS	0.01126606
C2B_THR	0.111749487	C1D_VAL	0.064134033	N7A_ASP	0.034717747	O5B_LYS	0.011256246
O7N_HIS	0.110573667	N1A_ASP	0.064125337	C5N_THR	0.034520477	C2A_LYS	0.011124762
C8A_SER	0.110288278	O7N_TYR	0.064086413	C3N_ASN	0.034431891	O2A_ARG	0.01061547
O5B_VAL	0.109365284	PN_ASP	0.063669197	C7N_LYS	0.03412763	O3B_TRP	0.010596586
O3_ASP	0.108192747	C2D_ALA	0.063214388	C2D_ARG	0.034126311	N1N_PHE	0.00998952
C1B_LYS	0.108082034	O3_MET	0.062869632	N1N_TYR	0.033524849	C7N_ALA	0.00994264
C4D_LEU	0.107729603	O1A_TRP	0.062853765	C5N_TRP	0.033375128	O2B_ASP	0.009908968
C1D_THR	0.107220378	C4D_ASP	0.062836247	C1D_SER	0.032376306	N7N_TRP	0.009672741

N7N_GLN	0.107145492	O7N_PHE	0.06270243	C5B_MET	0.032376033	N9A_SER	0.009415228
C2A_ALA	0.105908542	C5B_THR	0.062417335	C6A_GLN	0.031523082	N1A_LEU	0.009285992
C3N_THR	0.105248594	C6N_MET	0.062197946	C2N_ARG	0.031276336	O5B_PRO	0.009048099
C6A_ASN	0.104591342	PN_TRP	0.061994385	PN_HIS	0.030665142	C3D_ARG	0.00856326
C5N_ILE	0.104010781	C2D_VAL	0.061925295	O2N_GLY	0.030550952	N1A_CYS	0.008427029
N7N_PRO	0.102646095	PN_LYS	0.061636747	O7N_PRO	0.030053576	C6A_LEU	0.008178712
C2N_PHE	0.10249854	C3B_CYS	0.061045003	O4D_TRP	0.029930128	C4N_GLN	0.008101571
N3A_ARG	0.101773524	O1A_MET	0.060950449	C3B_GLY	0.029781147	C4A_VAL	0.008086476
C2A_THR	0.101567253	C8A_ARG	0.060817792	N1A_ARG	0.029752999	O5D_GLY	0.007969447
N6A_GLN	0.101434154	PA_CYS	0.060276098	C7N_GLY	0.029649647	C3B_ASN	0.007766154
N1N_GLN	0.100911738	C6N_TRP	0.059803838	O2B_HIS	0.029521975	C2B_TRP	0.0072598
C2D_GLU	0.100833855	O1A_ASN	0.059279603	C3N_ILE	0.028964874	N6A_ASN	0.007186107
C2N_ASN	0.099646904	PA_PHE	0.058835203	O2N_TRP	0.028578168	C2A_ARG	0.006983245
C2B_CYS	0.098308531	O5D_PRO	0.057889556	O7N_ASN	0.028227782	C1D_PHE	0.006853563
C2N_THR	0.097889728	C4N_PRO	0.057582026	N6A_LYS	0.027873306	N7N_LEU	0.006283492
N6A_ASP	0.097880648	O4B_CYS	0.05742591	O2N_LYS	0.027710043	C5D_GLN	0.00605615
C3D_TYR	0.097789016	C4D_SER	0.057417109	C3D_LYS	0.027588124	C3N_PRO	0.005892243
C3D_ASN	0.096020327	N9A_PRO	0.056977578	N1A_GLN	0.027196561	C7N_VAL	0.005799317
N7N_ARG	0.095925003	O4B_LEU	0.056890567	N6A_VAL	0.027155871	O1A_GLY	0.00542776
PA_GLU	0.095654393	C3N_VAL	0.056484245	C2N_GLY	0.027000196	O4B_ASN	0.005285506
C2A_HIS	0.094314915	O7N_GLU	0.055914303	C7N_HIS	0.026998926	N1A_TRP	0.004883473
O4D_VAL	0.093174086	O5D_ARG	0.055821931	O7N_LYS	0.026956111	C4D_GLN	0.004734092
O3D_PRO	0.092015331	O3B_CYS	0.054819911	C3B_SER	0.026905237	N9A_ARG	0.004626553
O7N_VAL	0.091513198	C2B_TYR	0.054126689	O3D_ILE	0.026902842	N3A_GLN	0.003542283
C4N_CYS	0.089453556	C7N_MET	0.053802096	N1N_ALA	0.02674033	O3D_SER	0.003501124
C5N_ALA	0.088879631	C1B_ASN	0.05365572	PN_ARG	0.026713488	C3D_GLY	0.003266954
C1D_LEU	0.088099191	O3D_PHE	0.053469827	C3B_GLN	0.02669302	O5D_LYS	0.003223613
O4D_GLY	0.086767653	C3D_LEU	0.053229595	C5N_CYS	0.026574053	C5A_CYS	0.003056985
N7N_MET	0.086633406	N3A_MET	0.053096854	C4D_HIS	0.02642775	C6A_HIS	0.003020458
O2N_VAL	0.085984478	C7N_GLU	0.05280589	O2B_CYS	0.025727259	C5D_GLU	0.00274718
C4A_PRO	0.085348112	PA_LEU	0.0527438	C4B_CYS	0.025414997	O5B_ASN	0.002626459
O7N_ILE	0.084973072	O5B_ALA	0.052622602	O3_SER	0.02531283	O2N_MET	0.00250402
C6N_PHE	0.084854801	O3B_ARG	0.052168269	C4A_HIS	0.025217149	C2D_ASN	0.002268232
O2A_MET	0.084659482	O5D_LEU	0.052144021	O2D_GLN	0.025168633	O3_LYS	0.002168882
N7A_GLY	0.084155597	O3D_CYS	0.051815901	O1A_HIS	0.024616979	N1A_HIS	0.001854708
O2D_GLU	0.084149479	C4A_SER	0.051093903	C4D_PRO	0.024612369	O1A_PHE	0.001652292
C2A_ASN	0.083940647	PA_MET	0.051091304	C6N_HIS	0.024028726	N1N_THR	0.001652148
N7N_TYR	0.083794294	C3D_VAL	0.05093086	C6N_ASP	0.023679437	C2N_TYR	0.001482017
PA_ASP	0.083171344	N3A_PHE	0.050731905	N7N_PHE	0.023380668	O5B_CYS	0.001317108
C4D_CYS	0.083049838	O2D_TRP	0.050600777	O5B_GLU	0.023362986	C5B_SER	0.001232059
C1B_CYS	0.082678811	C2B_HIS	0.050513385	PN_MET	0.023060629	N9A_CYS	0.001219865
O2B_THR	0.082490082	C6N_GLN	0.050409382	O1N_GLN	0.022780128	C5B_GLN	0.001179896
C3N_LEU	0.081724665	N7A_SER	0.049681009	N6A_LEU	0.022696971	O2A_TRP	0.000815129
N1N_HIS	0.081721236	C5N_ARG	0.04940221	C2B_GLN	0.022212051	C5A_ILE	0.000680242
O3_VAL	0.080908382	C3B_LEU	0.049214786	C2N_SER	0.021588548	C5D_ASP	0.000586944
C3N_GLN	0.080883263	C1D_PRO	0.049053943	N1N_TRP	0.021172989	C6N_GLY	0.000356168
N1N_ILE	0.080273148	C4N_THR	0.0488634	C3B_ARG	0.021026055	PA_ILE	0.0000763
		O1A_TYR	0.048695887	O2D_CYS	0.0206318		
NAD	weight	NAD	weight	NAD	weight	NAD	weight
C8A_ASP	0.236094258	C4B_VAL	0.084678643	C2A_GLY	0.04585673	PA_GLN	0.024167113
O4B_SER	0.214267146	C4N_GLU	0.084495222	C8A_CYS	0.045674108	N9A_HIS	0.023710343
C4B_ASP	0.211914031	C1B_PHE	0.082788143	C3D_ALA	0.045302776	N7A_LEU	0.023256866
C5B_ASP	0.211619835	O2B_PHE	0.082187984	C3B_MET	0.045277989	C8A_ILE	0.023221567
O5B_ASP	0.201497995	O3_ILE	0.081837268	PN_CYS	0.045070493	C3B_TYR	0.023129246
C5A_LYS	0.199153626	O4D_GLU	0.081724154	C7N_TYR	0.044696847	C2B_VAL	0.02300447
O2D_VAL	0.192728146	C4A_LYS	0.081269877	C4D_ARG	0.044656503	O2B_VAL	0.022694831
C2B_GLU	0.1793113	C4D_GLY	0.080939478	C5A_TRP	0.044048925	C5N_TYR	0.022381938
N6A_ALA	0.178674412	C6A_GLU	0.080923251	N6A_TRP	0.044048925	C4N_TRP	0.022043873
C5B_TRP	0.177316428	C3N_TYR	0.080480348	C1D_ILE	0.043144806	O2N_ASN	0.021764038
O1N_LEU	0.176228984	C1D_CYS	0.080016833	O2D_HIS	0.042941727	C2A_SER	0.021614551
C2B_GLY	0.174419789	O4B_PHE	0.079708441	PA_ARG	0.042894018	N6A_GLY	0.02146111
N7A_PHE	0.16528297	O2D_SER	0.07945322	C5A_VAL	0.042803352	C3D_CYS	0.02135515
O2N_LEU	0.161914369	O1N_VAL	0.078945222	C2B_MET	0.042765752	O5D_GLU	0.021243463
N7N_ASP	0.156998091	C5D_SER	0.078576218	C3B_ASP	0.042701922	N7A_VAL	0.021039061
O2N_PHE	0.156590241	C6A_LYS	0.077817193	C4A_CYS	0.042634678	N1A_TYR	0.020921701
O3B_PHE	0.155184371	C8A_ASN	0.077790099	PN_ALA	0.042627867	O1N_THR	0.020897222
C5B_ALA	0.152835734	N3A_LEU	0.077742066	C5A_ASP	0.042546411	N6A_CYS	0.020775768

O3B_GLY	0.151913164	N6A_PRO	0.077616001	C2N_GLU	0.042497144	O3_GLU	0.020196977
O2B_GLU	0.151898457	C2N_ALA	0.077249401	O5B_GLN	0.042340288	C5N_LEU	0.02009685
O1N_TYR	0.151450549	C7N_CYS	0.076845666	C1D_ASP	0.041495819	C4B_GLN	0.019926272
N3A_TYR	0.150626291	O3D_GLY	0.076306045	N9A_TRP	0.04128563	C1B_GLN	0.019633755
N9A_ASP	0.149611431	C3D_GLN	0.075904853	C2A_TRP	0.041271331	N7N_GLY	0.019561879
O5D_GLN	0.148609502	C5D_CYS	0.075202986	N3A_TRP	0.041271331	C4B_ILE	0.019405753
C2B_PRO	0.147532603	C6N_GLU	0.074741328	C2B_ALA	0.041255388	C7N_ARG	0.019372093
C4B_ARG	0.143967247	C5A_THR	0.074527957	N1A_THR	0.041026461	O1A_ASP	0.019244841
N3A_ASP	0.14300716	C2N_ASP	0.074459978	C2N_LEU	0.040888207	O2D_LYS	0.019194013
O2B_LEU	0.141964964	O2A_GLY	0.073382866	O5D_VAL	0.040156633	O1N_PHE	0.019158342
O3B_GLU	0.141180163	O4B_HIS	0.073219477	O3B_GLN	0.039927468	C5D_THR	0.019113421
N9A_LEU	0.139362426	C6N_LYS	0.072117998	O3D_ALA	0.039896181	C8A_GLN	0.018997788
C1B_ALA	0.136457677	C1B_TRP	0.070639586	N1N_GLY	0.039653791	C3N_ARG	0.018983506
C4D_TYR	0.13469436	C5A_ALA	0.069382114	O3B_HIS	0.039458724	O4D_SER	0.018702347
O1N_ARG	0.133895846	O2D_THR	0.069185612	C3D_ILE	0.039401154	C3N_ASP	0.018453586
C3N_GLU	0.127710335	C3B_VAL	0.069108754	N6A_ARG	0.039362501	C6A_CYS	0.018355811
O5D_ALA	0.127675979	C5N_ASN	0.069004812	O1N_CYS	0.039281688	O3D_GLN	0.018256316
N7N_ILE	0.126107558	C5N_GLY	0.068747065	N6A_SER	0.03921673	O2N_GLU	0.01817712
N9A_GLU	0.125364201	C2B_ILE	0.0686374	C2D_MET	0.03895439	C1D_GLN	0.017972709
O1A_ARG	0.124914545	O2D_LEU	0.068182443	C5D_VAL	0.038878843	C3B_HIS	0.017919884
O3B_ILE	0.124643569	C6N_LEU	0.068135623	C5D_LEU	0.038872236	C4N_PHE	0.017696368
C8A_MET	0.122164263	O1N_GLU	0.068107524	C1B_MET	0.038286853	C2D_TRP	0.01744367
O5D_THR	0.122063447	C5B_ARG	0.068011606	O4D_PRO	0.038198556	C2D_CYS	0.016935276
O1N_ASN	0.121086105	C4D_TRP	0.067954707	O1N_PRO	0.03813499	O4D_LYS	0.016476418
O7N_ARG	0.119867381	O4D_ASP	0.066602311	N1N_SER	0.038079186	C4B_TRP	0.016383175
N3A_LYS	0.119833863	C7N_THR	0.0658177	C6A_PHE	0.038067227	C7N_ASP	0.016178585
N1A_PHE	0.119547183	C3N_HIS	0.065214102	C5D_GLY	0.038048449	PA_LYS	0.016159851
N9A_ILE	0.119507247	C8A_THR	0.065129865	C5B_ASN	0.038045075	C2A_PRO	0.016101081
C3B_ILE	0.117697785	O7N_LEU	0.064726578	C6A_ASP	0.037830366	C5N_VAL	0.016097103
C1B_SER	0.117485882	C4D_PHE	0.064478493	N7A_LYS	0.037771646	N3A_SER	0.015498109
C5A_PRO	0.116871565	C1D_ALA	0.064318425	O1A_THR	0.037593202	C3D_SER	0.015402666
C3D_HIS	0.116540224	C5D_LYS	0.063959291	C4A_ARG	0.037365628	N1N_MET	0.015128158
C8A_HIS	0.115036079	C4N_LYS	0.063207444	C6A_TRP	0.037202519	C5B_PRO	0.015098312
C1B_GLU	0.114836695	O3D_MET	0.062607471	O7N_TRP	0.037119054	C2D_LEU	0.014951664
O4B_PRO	0.114719939	O5B_LEU	0.061997954	O3D_VAL	0.036711513	O2A_CYS	0.014949863
C5A_GLY	0.114416309	O2A_ASN	0.061732721	C8A_LEU	0.036654066	O3D_GLU	0.014843767
C2A_ASP	0.113940715	N6A_THR	0.061577713	N1N_GLU	0.036633975	O2N_SER	0.014469964
C3N_ALA	0.113621507	C7N_ILE	0.061382838	C4A_ILE	0.036210335	O2B_MET	0.014381374
PN_ASN	0.113168012	O3_GLN	0.061367368	C4A_TRP	0.036192205	C3D_GLU	0.013753582
N6A_HIS	0.113024505	C2N_HIS	0.061266738	C2N_GLN	0.035567033	O2A_PRO	0.013548182
C5A_MET	0.112329829	C2N_PRO	0.060988083	C3B_TRP	0.034880319	N7A_GLN	0.012765498
PA_GLY	0.11165329	C6N_THR	0.060959542	O5B_HIS	0.03477746	C3D_TRP	0.012640582
O2A_GLU	0.111244424	PA_THR	0.060328547	C1D_MET	0.0347337	O1N_ALA	0.012390965
O4B_GLU	0.109619937	C6N_TYR	0.060307835	C5N_ASP	0.034172545	O4B_THR	0.012108389
C4D_ALA	0.108202492	N7N_VAL	0.060165214	O4B_ALA	0.034119062	O4D_GLN	0.012038157
O2B_ASN	0.108147068	O5D_CYS	0.059410492	C4A_GLY	0.033689269	N1N_ASN	0.011838299
O2D_ARG	0.108038157	C3N_PHE	0.059088755	O4D_ALA	0.033586427	N9A_LYS	0.011775881
C7N_ASN	0.107213425	O2N_CYS	0.059056218	O1N_GLY	0.033556184	C2D_GLY	0.011721179
O1N_ILE	0.106640874	N1A_ASN	0.058117727	C4B_GLY	0.033344892	O7N_GLN	0.011160851
C2A_TYR	0.106180162	PN_GLY	0.057992715	O3_PHE	0.033049727	O1N_MET	0.011066817
C8A_PHE	0.105426384	O5D_ASN	0.057907553	C3B_THR	0.032956614	C1B_HIS	0.010953044
N3A_THR	0.104176435	O2N_ARG	0.057788184	C3B_PRO	0.032916451	C3B_ALA	0.010883357
N1N_PRO	0.102629164	C4D_VAL	0.057773452	C5B_GLU	0.032826056	PN_TYR	0.010703707
O5D_SER	0.10245746	C2N_CYS	0.057688219	C7N_TRP	0.032662711	N9A_ALA	0.010443668
C4B_GLU	0.102381935	N9A_VAL	0.057370037	C5N_GLN	0.032605112	C4B_ASN	0.010442276
O7N_ALA	0.102269399	C1B_ILE	0.05698411	O4B_TRP	0.032570682	N9A_GLN	0.009324513
C5B_LYS	0.102210074	O2B_ALA	0.056981668	C4B_PHE	0.032417942	N7A_GLU	0.00883231
C2A_GLN	0.101828872	C4N_TYR	0.056651445	C5B_HIS	0.03236729	C5A_PHE	0.008825354
O1A_GLN	0.101685897	C3D_ASP	0.056555527	N1A_GLY	0.03232162	O3B_LEU	0.008819554
C8A_GLU	0.100650034	C6A_PRO	0.056529942	N3A_CYS	0.032317815	C6N_PRO	0.008639054
N7N_THR	0.099880328	C1D_HIS	0.056345666	C1B_PRO	0.031301457	N1N_LEU	0.008632498
O7N_THR	0.099664754	O3D_ASN	0.056314177	O2B_PRO	0.03121384	C5D_HIS	0.007959814
C4N_ALA	0.098701899	C4B_HIS	0.056175761	O4D_ASN	0.030787145	C2D_ASP	0.00768146
O3D_THR	0.097248115	C4A_ASP	0.055724142	N7A_TRP	0.030680017	O1A_LEU	0.007670774
O4B_ASP	0.096754876	C8A_TRP	0.055617695	PN_LEU	0.029719512	C5A_HIS	0.007650503
C8A_LYS	0.095315866	C5N_PRO	0.055392821	O2A_GLN	0.029703758	O3B_TYR	0.007502293
C1B_ASP	0.094624659	O3_CYS	0.055226603	O5D_MET	0.029487202	C2D_GLN	0.007488351
N6A_GLU	0.094601341	N6A_PHE	0.054991735	PA_TRP	0.029056729	C2A_LEU	0.007471721

N7N_CYS	0.094435354	C4A_GLU	0.05479858	O3_THR	0.029037937	O3_ARG	0.006893838
N3A_GLU	0.094044915	C2B_LEU	0.053824457	O2D_ILE	0.028969805	PN_THR	0.006589932
C2N_MET	0.093635101	O5B_ARG	0.05370051	C2D_PHE	0.028951052	C6N_ILE	0.006317509
C7N_SER	0.093273592	C3N_TRP	0.053645898	C2D_THR	0.028863737	O3B_SER	0.00628109
N7A_PRO	0.092887422	C5N_LYS	0.053172349	C6A_THR	0.028566318	N1N_CYS	0.006183564
C2A_GLU	0.092512112	O4B_GLN	0.053113516	C7N_PHE	0.0284569	O2N_HIS	0.006143172
C8A_VAL	0.092039322	N1A_VAL	0.052949956	C2D_HIS	0.028138399	C4B_ALA	0.005844543
C2D_PRO	0.091974416	N7A_ILE	0.052668712	O2A_SER	0.027907848	O2A_ASP	0.005816768
C5A_ARG	0.091774884	C2B_ASN	0.052129331	C8A_PRO	0.027839496	O2N_ILE	0.00570703
O7N_GLY	0.091607528	O2A_LYS	0.052029338	C2D_LYS	0.027755749	N3A_ILE	0.005557604
C4A_PHE	0.091478957	O5B_TRP	0.051926206	C5B_VAL	0.027696387	C4D_GLU	0.005447283
N7A_HIS	0.091434955	PN_ILE	0.051876557	C4N_VAL	0.027305555	O1N_ASP	0.005178466
C3B_GLU	0.09139035	C3N_CYS	0.050868751	O2N_ALA	0.027160743	O3B_VAL	0.004899055
C4A_LEU	0.090872072	C4B_MET	0.050634	C7N_LEU	0.027019403	C4N_ASN	0.003532887
O3B_ASP	0.090720383	C6A_ILE	0.050325083	O2D_ASN	0.026645431	C3D_PHE	0.003504186
C4B_PRO	0.09067722	O1N_SER	0.049049516	O3_LEU	0.026576686	C6N_VAL	0.003094678
N7A_ARG	0.090528536	O7N_ASP	0.048624583	O5D_PHE	0.026565767	O5B_TYR	0.002555715
N7A_MET	0.090306467	N6A_MET	0.048577237	PA_HIS	0.026455381	C3N_MET	0.002362533
C1B_LEU	0.089465086	O4D_ARG	0.048423773	O5B_THR	0.026232191	C5N_GLU	0.002179866
C5A_GLU	0.088046164	O2D_TYR	0.048363667	C6A_MET	0.026165631	O3B_THR	0.001955582
N7A_THR	0.088005608	PA_PRO	0.048103131	C4B_LYS	0.026110914	O1A_PRO	0.001622565
N7N_GLU	0.087465764	C6N_CYS	0.047847513	C5D_MET	0.025805668	C1D_GLY	0.001541014
N3A_ALA	0.087417513	C6N_SER	0.047755285	N1A_LYS	0.025796286	C4A_TYR	0.001506384
N9A_MET	0.087399922	O2D_MET	0.047508299	C5N_PHE	0.025618651	O4B_ILE	0.000794275
N9A_ASN	0.08732569	O2A_ILE	0.047073498	N1A_MET	0.025546119	N7A_CYS	0.00025527
O2N_TYR	0.087107984	PA_SER	0.046824837	C4A_THR	0.02553769	O5D_ASP	0.000240182
N9A_PHE	0.086900914	N7N_HIS	0.04665957	N1A_GLU	0.025466655	O4D_LEU	0.000196388
O3_HIS	0.08674575	C1B_VAL	0.04659479	C2A_ILE	0.025021287	C5D_PHE	0.000107303
O4D_ILE	0.086687568	C7N_PRO	0.04631959	C4A_MET	0.024798013	PN_SER	0.000103847
O2N_THR	0.086066019	O4B_MET	0.046008828	C4N_ILE	0.024773603	O1A_VAL	0.0000672
C6A_ALA	0.085894582	N6A_TYR	0.045938949	O4B_LYS	0.024320924	N1N_VAL	0.0000641
C2B_ASP	0.085308204						