UiT The Arctic University of Norway

Faculty of Biosciences, Fisheries and Economics

Development of a bioinformatic framework for the phylogenetic and structural analyses of protein evolution and co-evolution

Yin-Chen Hsieh

A dissertation for the degree of Philosophiae Doctor – May 2024

Development of a bioinformatic framework for the phylogenetic and structural analyses of protein evolution and co-evolution

Yin-Chen Hsieh

A dissertation for the degree of Philosophiae Doctor

May 2024

UiT The Arctic University of Norway

Faculty of Biosciences, Fisheries and Economics

Department of Arctic and Marine Biology

Microorganisms and Plants Research Group

Acknowledgments

There were many people who supported me in this thesis journey, and therefore a lot of thank yous to be said. I hope to be able to express this gratitude in person to you directly, but here is the general gist in writing, which by no means can do justice to just how much I appreciate you and what you have done for me.

Thank you Ines, for supervising, guiding, and advocating for me not only through science but through this chapter of my life, in the wisdom you shared and the numerous hikes we did, and for including me in your farm and Tromsø life overall. Not every PhD student gets their ass kicked regularly by their supervisor's rooster, but maybe everyone should - it is a very humbling experience.

Thank you Alex S, for co-supervising me from afar, for your ideas and intense enthusiasm, and especially your unfailing energy.

Thank you as well to all my co-authors for their contributions to this work, especially Alex H, for being the mTOR expert and for the good times exploring Innsbruck together.

I also want to thank everyone in the Thedieck and Ziegler groups, for the fruitful collaborations, the numerous times you hosted me, and for the fun experiences shared in the lab, on boats, eating strange fish, and on snowy mountains.

Thank you Selina for doing your Masters thesis with me, and for showing me just what structured organisation can bring about. You taught me more than I taught you!

Thank you to the colleagues in the MicroPlants group, who made the workplace very inviting. I ended up never winning this wine lottery thing, but the cakes were excellent!

Then to my friends, thank you so much. You were my Tromsø (and 'Southern' Europe) family, and provided me with all types of encouragement, advice, and fun along the way. This is directed at, in no particular order: Jørn, Lena, Kathi, Mathias, Ligia, Roland, Suraj, Stijn, Linn, Oskar, Marit, Philo, Julia M, Julia W, Ollie, David, and Nate. You've truly shown me what it is like to be in good company, and to be well taken care of. I hope to continually return the favor in the years to come.

Thank you to my parents and brother, for supporting me from another continent, and for all the goofing off and laughs we had (almost) every Sunday.

I have the feeling I missed some names, that this page should be several times longer, and that I will remember only after this thesis is printed. A thank you in advance to those who I will only thank in person!

Summary

This thesis investigates the evolution and co-evolution of proteins and their implications in molecular biology through a bioinformatics approach. It develops a framework for studying these processes that operates across the taxonomic scales and at the molecular level. This framework generates a large-scale overview of evolutionary patterns and pinpoints specific interaction sites in proteins, leveraging state-of-the-art sequence co-evolution and protein structural prediction techniques. More specifically, the framework constructs two tools: a tool for the phylogenetic mapping of protein presence-absence, and a pipeline incorporating Direct Coupling Analysis for detecting inter-protein co-evolution, and develops strategies for use of AlphaFold2 for predicting complex protein interactions. Utilizing this framework, this research addresses critical questions in molecular biology, particularly in the context of pathway evolution, organism-specific pathway preferences, and protein-protein interactions. The studies presented in this thesis demonstrate the application of these tools in various biological contexts involving enzymes in the NAD recycling pathway and regulatory proteins in the mTOR signaling network. The developed framework not only advances our understanding of molecular biology but also opens up new avenues for experimental validation of predicted interactions and further computational exploration, especially towards integration of various types of results. Future improvements to this framework will focus on incorporations of sequence-based information on paralogues and post-translational modifications, improving the interpretability of results, and incorporating advanced structural prediction models to better handle the genomic complexity of eukaryotic systems. This thesis underscores the potential of bioinformatics in transforming our approach to studying molecular evolution and sets the stage for future research that could further refine these methodologies and expand their applicability.

Sammendrag

Denne avhandlingen utforsker evolusjonen og samevolusjonen av proteiner og deres implikasjoner i molekylærbiologi gjennom et bioinformatisk perspektiv. Den utvikler et rammeverk for å studere disse prosessene som opererer på tvers av taksonomiske skalaer og på molekylært nivå. Dette rammeverket gir en oversikt i stor skala over evolusjonsmønstre og identifiserer spesifikke interaksjonssteder i proteiner, ved å utnytte toppmoderne teknikker for sekvens-sameevolusjon og proteinstrukturprediksjon. Mer spesifikt konstruerer rammeverket to verktøy: et verktøy for fylogenetisk kartlegging av protein tilstedeværelse-fravær, og en pipeline som inkorporerer Direct Coupling Analysis for å oppdage interprotein koevolusjon, samt utviklede strategier for bruk av AlphaFold2 for å forutsi komplekse proteininteraksjoner. Ved å bruke dette rammeverket adresserer denne forskningen kritiske spørsmål i molekylærbiologi, spesielt i konteksten av biologisk vei(pathway)-evolusjon, organisme-spesifikke vei-valg og protein-protein interaksjoner. Studiene presentert i denne avhandlingen demonstrerer anvendelsen av disse verktøyene i ulike biologiske sammenhenger som involverer enzymer i NAD-resirkuleringsveien og regulatoriske proteiner i mTOR-signaleringsnettverket. Det utviklede rammeverket fremmer ikke bare vår forståelse av molekylærbiologi, men åpner også nye veier for eksperimentell validering av forutsagte interaksjoner og videre datamaskinbasert utforskning, spesielt mot integrering av ulike typer resultater. Fremtidige forbedringer av dette rammeverket vil fokusere på inkorporering av sekvensbasert informasjon om paraloger og post-translasjonelle modifikasjoner, forbedring av tolkbarheten av resultater, og integrering av avanserte strukturprediksjonsmodeller for bedre å håndtere den genomiske kompleksiteten til eukaryote systemer. Denne avhandlingen understreker potensialet for bioinformatikk til å transformere vår tilnærming til å studere molekylær evolusjon og legger grunnlaget for fremtidig forskning som kan videre utbedre disse metodene og utvide deres anvendbarhet.

List of papers and declaration of contributions

Paper 1:

Yin-Chen Hsieh, Mathias Bockwoldt, Ines Heiland.

VisProPhyl - Visualisation and analysis of protein phylogenetic presence-absence.

Manuscript research article under review at BMC Bioinformatics

Paper 2:

Suraj Sharma, Yin-Chen Hsieh, Jørn Dietze, Mathias Bockwoldt, Øyvind Strømland, Mathias Ziegler, Ines Heiland.

Early Evolutionary Selection of NAD Biosynthesis Pathway in Bacteria.

Research article published in MDPI Metabolites 2022 (DOI: 10.3390/metabo12070569)

Paper 3:

Yin-Chen Hsieh, Alexander Heberle, Kathrin Thedieck, Ines Heiland.

Structure prediction in the mTOR-verse: evaluation of the predictive performance of AlphaFold Multimer on the G3BP:TSC complex and 4EBP:eIF4E complex in mTOR signaling.

Manuscript research article prepared for submission to Biophysical Journal

	Paper 1	**Paper 2**	**Paper 3**
Concept and idea	MB, IH	MZ, IH	YH, AH, KT, IH
Study design and methods	YH, MB, IH	SS, YH, JD, MB, MZ, YH	YH, AH, KT, IH
Data gathering and interpretation	YH, IH	SS, YH, ØS	YH, AH, KT, IH
Manuscript preparation	YH, MB, IH	SS, YH, JD, MB, ØS, MZ, IH	YH, AH, KT, IH

Table 1: **Contributions to Papers 1, 2, and 3**: More specifically, in Paper 2, YH designed, conducted, and wrote the manuscript for the phylogenetic analysis, and SS did the same for the metabolic modeling, which is not factoring into this thesis.

List of figures

1 **Illustration of covariation in MSA:** How co-evolutionary methods work: reciprocal mutations, or covariation, detected between positions *a* and *b* are given a score according to a global statistical model that calculates coupling scores across all pairwise combinations of residues in this inter-protein case. MSAs corresponding to the proteins are labeled with the target sequence in a colored box, with the color matching the structure. Structures were rendered via PyMOL Schrödinger, LLC [2015]. 7

2 **Workflow Diagram for Phylogenetic Presence-Absence Analysis:** All steps correspond to the *phylogenetics.py* script in the GitHub repository: `https://github.com/MolecularBioinformatics/VisProPhyl`. The left-most column shows the main procedure of sequence file querying, sequence hit filtering, and result combination, followed by heatmap and tree visualisations. Arrows to the right show the file outputs per step, and their location in the folder scheme. Refer to Paper 1 for an in-depth methodological explanation of the workflow. 16

3 **Workflow Diagram for Co-evolutionary Analysis:** All steps correspond to the *run_workflow.py* script in the GitHub repository: `https://github.com/MolecularBioinformatics/euk_dim_dca.git`. The left-most column shows steps from preliminary input file preperation through the start to the end, passing through five colored stages for a total of 10 steps in total. The general flow follows target sequence identification, sequence querying to identify homologs, extraction of reference sequences for the homologs, filtering of homolog sets based on common organisms across the two sets, multiple sequence alignment of homologs to the target, and co-evolutionary analysis. The middle column details output file formats, and the right-most column gives a concise technical description of each step. 17

4 **Visualisation Plots of Co-evolutionary Scores:** Hairball network visualizations are utilized to depict linkage patterns among interprotein residue pairs that score within a specified range of co-evolutionary values. Nodes are amino acids, node sizes represent the degree of each node, and nodes are colored based on the protein they originate from. These visualizations are generated as bipartite graphs. Histograms illustrate the distribution of DCA scores and conservation. Additionally, parallel plots highlight highly co-evolving residues between two proteins, with each protein's residue indices displayed on separate vertical axes. Lines indicate detected co-evolution, with line color indicating the magnitude of the score. 20

5 **Co-evolutionary Plots for Raptor and 4EBP1**: Comparison of top 100 scoring residue pairs (parallel plot, left panel) with top 400 scoring residue pairs (hairball plot, right panel) shows no co-evolutionary signal between the 4EBP1 TOS with the Raptor binding pocket. Sections of the Raptor binding pocket are weakly co-evolving with C-ter residues in 4EBP1 in front of the TOS, in residues 104-108, as shown in the hairball plot, bottom right circle. 30

6 **Co-evolutionary Plots for Raptor and S6K** Consistency between the parallel plot (top 100 scoring residue pairs) with the hairball network plot (top 400 scoring residue pairs) strongly suggests residues of the S6K TOS region coevolving with one residue in the Raptor binding pocket. 30

7 **4EBP1:Raptor Parallel Plot with Conservation Patterns**: The parallel plot of top 100 scoring residue pairs (same parallel plot as in Fig. 5) is juxtaposed with two conservation histograms of 4EPB1 (left) and Raptor (right). Co-evolutionary signal for 4EBP1 tends to come more out of regions of higher variability, but the TOS region itself (residues 114-118) does not have high variability. Raptor shows a lengthy conserved stretch that covers the binding pocket, with more variability from residues 700 onwards, with peaks around residue 900. 31

8 **S6K:Raptor Parallel Plot with Conservation Patterns**: Visualisation of the S6K:Raptor top 100 scoring residue pairs (same parallel plot as in Fig. 6) together with conservation scores of both proteins reveals high variability in the N-ter region of S6K (containing the TOS region) that is also found to highly coevolve with residue 396 in Raptor, which is itself variable. S6K has another highly variable region around residue 400, but no clear co-evolving signal with Raptor is found out of this region. 32

9 **G3BP1:TSC2 Hairball Network Visualisations**: Comparison of the top 100 scoring residue pairs between G3BP1 and TSC2 with the top 400 scoring residue pairs identifies two clusters around G3BP1 node/residue 325 and 172. These two clusters grow to be highly connected with a strongly co-evolving neighborhood of residue pairs in the top 400 hairball plot. 33

10 **G3BP1:TSC2 Parallel Plot with Conservation Patterns**: The parallel plot for the top 100 scoring residue pairs are accompanied by conservation score histograms for G3BP1 and TSC2 highlights two prominent regions from the TSC2 perspective: residues 980-1010 and another region encompassing residues 1760-1780. Conservation scores across both proteins reveal no clear association between higher variability and stronger co-evolutionary signals. 33

11 **MSA visualisation and evaluation component**: Proposed addition to the dual-axis framework, created over each input MSA. MSA sequences are clustered based on a consistent sequence similarity threshold, to allow for comparison of MSA depth. Sequences are visualised in a circular tree, and highlighted based on taxonomic group of interest. 38

List of abbreviations

4EBP1/2	eukaryotic initiation factor 4E-binding protein 1 and 2
AF2	AlphaFold2
F2M	Alphafold Multimer
APT	adenine phosphoribosyltransferase
CPU	central processing unit
DCA	direct coupling analysis
DNA	deoxyribonucleic acid
eIF4E	eukaryotic initiation factor 4E
G3BP1/2	Ras GTPase-activating protein binding protein 1 and 2
GPU	graphics processing unit
HPC	high-performance computer
HPRT	hypoxanthine-guanine phosphoribosyltransferase
IDR	intrinsically disordered regions
MSA	multiple sequence alignment
MI	mutual information
mTOR	mechanistic target of rapamycin
mTORC1/2	mTOR complex 1 and 2
NAD	nicotinamide adenine dinucleotide
NAAD	nicotinic acid adenine dinucleotide
NAMN	nicotinic acid mononucleotide
NAMPT	nicotinamide phohsphoribosyltransferase
NA	nicotinic acid
Nam	nicotinamide
NAPRT	nicotinate phosphoribosyltransferase
NCBI	National Center for Biotechnology Information
NRIS	Norwegian research infrastructure services
Neff	effective number of sequences
NMN	Nam adenine dinucleotide
NMNAT	nicotinic acid mononucleotide transferase
nr	nonredundant protein database
NN	neural network
PncA	nicotinamidase (bacterial name)
PPI	protein protein interactions
PRT	phosphoribosyltransferase
PTM	post-translational modifications
QA	quinolinic acid
S6K	S6 kinase
TSC	tuberous sclerosis complex
UPRT	uracil phosphoribosyltransferase

Contents

Acknowledgments **i**

Summary **ii**

Sammendrag **iii**

List of papers and declaration of contributions **iv**

List of figures **v**

List of abbreviations **vii**

Contents **viii**

1 Introduction **1**

1.1 The evolutionary journey: from Darwin's theory to modern molecular evolution . 1

1.2 Unraveling co-evolution: insights into reciprocal evolutionary pressures and molecular interactions . . . 3

1.3 Computational advances in detecting protein co-evolution: from mirrortree to direct coupling analysis . . . 5

1.4 Advancements in protein structure prediction: the rise of neural network methods 8

1.5 Test case 1: NAD pathway from protein perspective . . . 8

1.6 Test case 2: mTOR signaling from a protein perspective . . . 10

1.7 Thesis objectives . . . 12

2 Materials and methods **14**

2.1 Data . . . 14

2.2 Phylogenetic presence-absence analysis . . . 15

2.3 Protein sequence co-evolutionary workflow . . . 15

2.3.1 Stages 1 - 3 . . . 16

2.3.2 Stages 4 - 5 . . . 18

2.3.3 Post-processing: Visualisation of co-evolutionary scores and MSA analysis 19

2.4 Protein structure and PPI prediction . . . 21

2.5 Computing platforms and HPC job submissions . . . 21

3 Results **23**

3.1 Summary of the papers . . . 23

3.1.1 Paper 1. VisProPhyl - Visualisation and analysis of protein phylogenetic presence-absence. . . . 23

3.1.2 Paper 2: Early Evolutionary Selection of NAD Biosynthesis Pathway in Bacteria. . . . 25

3.1.3 Paper 3: Structure prediction in the mTOR-verse: evaluation of the predictive performance of AlphaFold Multimer on the G3BP:TSC complex and 4EBP:eIF4E complex in mTOR signaling. 26

3.2 Unpublished results 28

3.2.1 Co-evolutionary detection in mTOR-associated proteins 29

4 Discussion **34**

4.1 Adopting a dual-axis framework: extracting co-evolutionary and evolutionary patterns from protein sequence data 34

4.2 Navigating opportunities and challenges in analysing eukaryotic protein sequence data 37

4.3 Understanding the interpretability of results in a broader context 41

5 Conclusion and outlook **43**

Bibliography **45**

6 Papers 1-3 **54**

1 Introduction

This thesis is fundamentally about co-evolution, the process in which two or more entities undergo evolutionary change as a result of reciprocal pressure between them, and how the search for clues of co-evolution on a molecular level has advanced the field of protein biology, from a bioinformatic perspective. The three papers and additional work constituting this thesis combine co-evolution and evolutionary theory into an informatics framework to investigate two critical types of questions in molecular biology: first, how to detect and use protein presence-absence to probe biological pathway evolution and organism-specific pathway preferences, and second, how to predict protein structure and protein-protein interactions for proteins of varying disorder, size, and complexity such that accurate, *de novo* structural predictions can be made.

The first work, Paper 1, presents a bioinformatic tool that gives a birds-eye-view of the current status in protein evolutionary and co-evolutionary patterns across the tree of life, in the form of a phylogenetic overview of protein presence-absence patterns in selected taxonomic groups of organisms. Paper 2 then utilises this tool to distinguish between bacterial groups based on their selective preference for recycling pathways of essential compounds, and thereby sheds light on the question of the extent to which co-evolution and evolution overall play a role in influencing this preference, particularly for bacterial groups associated with mammalian hosts. Paper 3 switches gears and dives deep into interprotein co-evolution by examining the ability of the state-of-art in protein structure prediction to use co-evolutionary signals in a neural network setup to predict protein-protein interaction sites, focusing particularly on its capabilities in tackling flexible, large, protein complexes. Additional, unpublished research concentrated on a pipeline designed for detecting inter-protein co-evolution, aiming to pinpoint interaction regions among eukaryotic proteins and explore potential signatures of co-evolutionary signals suggestive of interaction.

This introduction derives protein co-evolution from its molecular basis in evolution and translates these concepts into sequence-based information. Then, it sets the stage for using this information within phylogenetic pathway analysis and protein structure and interaction prediction. It presents the biological test cases, the NAD salvage pathway and mTOR signaling pathways, and concludes with the main objectives of this thesis.

1.1 The evolutionary journey: from Darwin's theory to modern molecular evolution

Co-evolution is derived from evolution; therefore, any discussion of these concepts must begin with a discussion of evolutionary theory, which is best understood in the context of its historical timeline.

Although widely accepted in the 21st century, the concept of evolution as we know it has itself evolved through time, surviving contention and undergoing numerous updates since Charles Darwin first presented the idea in his groundbreaking *Origin of Species* [Darwin, 1859] in 1859. Much of Darwin's original ideas still hold: that all living organisms stem from a common ancestor, that there exists natural variation within a population of organisms, and that differences between distinct groups of organisms or species observed today arise from inherited variation across generations of ancestral organisms. These ideas are collectively referred to as *descent by modification*, and are glued together by Darwin's theory of *natural selection*, whereby variation in the form of phenotypic traits is, over time, selected for and passed on when it confers a survival or reproductive advantage to an individual over others in the same population. Darwin's theory was constructed on the macroorganismal level, focusing on observable traits, as opposed to the molecular level, on which most of evolutionary work is done today. It was not until the independent work of Gregor Mendel (1865), Hugo de Vries (1900), and Correns and Tschermak (1900) [Allen, 2003] that Darwin's theory of evolution found its basis at the molecular level, in an inheritable unit known as a *gene.*

The discovery of genes kicked off a half-century of groundbreaking genetics research, where a string of monumental discoveries were made that pushed forward understanding of the mechanisms of evolution. Genetic variation in the form of mutations was characterised by de Vries, thereby providing a plausible means for genes to change, and the discovery of chromosomes by Flemming, van Beneden, and Morgan [Hamoir, 1992; Paweletz, 2001; Benson, 2001] provided the actual inheritable unit carrying the genes. The information contained in genes was decoded as deoxyribonucleic acid (DNA), chains of adenine, thymine, cytosine, and guanine linked together by a sugar-phosphate backbone, and the molecular language of genetic information was identified. In 1953, the work of Watson, Crick, Wilkins, and Franklin culminated into the A-T, C-G base-pairing paradigm and the structure of DNA [WATSON and CRICK, 1953; FRANKLIN and GOSLING, 1953]. Shortly thereafter, the universal genetic code, a triplet-codon mapping from DNA to amino acid, was deduced [Crick, 1968]. The brief summary of these achievements presented here potentially makes them sound commonplace, which would be incorrect - they were revolutionary. Together, these achievements effectively launched the field of molecular biology, and paved the way for the formation of the modern theory of molecular evolution.

The modern theory of molecular evolution is that evolution arises from propagation of variation in genes in the form of DNA mutations, which occur by random chance and also due to environmental factors. These mutations can be single point mutations or on the scale of entire gene duplications, which are common in more complex organisms such as the eukaryotes. Contrary to Darwin's theory of natural selection, however, the majority of DNA mutations that are inherited are thought to be neutral, in the sense that they do not represent a change in gene products and therefore do not impact their function, and would not be positively or negatively selected for. Neutral mutations are possible due to the degeneracy of the genetic code, where multiple triplet codons encode the same amino acid, as there is allowance usually in the third nucleotide of a codon for variation without changing the amino acid it encodes for. Such synonymous mutations are a large source of variation within a population, and their frequency is compared to non-synonymous mutations to discern the effects of selection. The generally accepted idea today is that natural selection as Darwin posited has less of a role in bringing about evolution as was originally assumed, and evolution is recognised as a result of a multitude of factors, some of which are: genetic drift, gene flow, non-random mating, and mutation-driven.

The idea that the majority of inherited variation is neutral was posited by Kimura (1968-1977) and King and Jukes (1969) as the neutral theory of molecular evolution [Kimura, 1968; King and Jukes, 1969]. They compared amino acid exchange rates to nucleotide exchange rates in highly conserved proteins and found the former to be extremely low compared to the latter. From a biological standpoint, this means that functional conservation is inbuilt into the system, amidst an environment of random variation that becomes fixated throughout generations, in a process termed *genetic drift*. Some genetic variation confers a slight advantage, others a slight disadvantage, but the majority brings about no change to the evolutionary status quo. Further comparative studies of the evolution of proteins and protein domains corroborate these findings, and indicate that higher variation accumulates in non-functional regions of proteins, whereas functional regions tend to be conserved. The neutral theory of molecular evolution has since been revised to a nearly-neutral theory, upon discovery of non-neutral effects arising from synonymous mutations [Shen et al., 2022], effects that impact gene expression, mRNA stability [Zhou et al., 2016], and numerous other downstream processes [Buhr et al., 2016]. The idea of molecular evolution is therefore still in flux, undergoing revisions based on new data. These findings, combined with current progress in sequencing and data availability of protein sequences, present a rich platform from which to study molecular evolution.

Focusing on protein evolution as opposed to nucleotide-level evolution is a more direct way to study the effects of evolution, as seen from a functional standpoint, since proteins carry out function in the biological context. This thesis therefore focuses on molecular evolution and co-evolution at the protein level.

1.2 Unraveling co-evolution: insights into reciprocal evolutionary pressures and molecular interactions

Co-evolutionary pressures arise from sustained interactions that select for certain traits within an ecological or molecular niche. Wherever evolution occurs, co-evolution most certainly occurs alongside it, since no species on this planet exists solely in a vacuum, and the scarcity of resources brings about intense interactions that, over time, select for reciprocal changes that drive co-evolution. This has been observed between many organisms, and a classic example is the co-evolution of pollinating machinery between the Malagasy orchid plant, *Angraecum sesquipedale*, and its pollinating hawkmoth, whereby the length of the orchid's nectar tube and the length of the moth tongue purportedly participated in a mutual elongation 'arms race' [Darwin, 1862] such that pollination and nectar retrieval could be optimised [Johnson and Anderson, 2010].

Co-evolution not only occurs between organisms but also inside individual organisms, namely between the cellular components that carry out their biological function. After all, like evolution, the basis of co-evolution is found at the genetic level, and the effects of it are compounded upwards, such that co-evolutionary changes ultimately manifest in the phenotype of an organism.

The concept of co-evolution, coined by Ehrlich and Raven in 1964 [Ehrlich and Raven, 1964], is actually as old as the concept of evolution [Darwin, 1862] and has similar roots in macrolevel observations, in the form of phenotypic traits that have evolved due to reciprocal pressure between interacting species. Abundant examples of co-evolving traits are found in nature, such as host-parasite co-evolution between plants and fungi that drives up plant resistance and pathogen virulence [Thrall and Burdon, 2003] and the symbiotic co-evolution of pollinating interfaces between plants and their pollinators [Darwin, 1862; Pauw et al., 2009], mentioned earlier. Co-evolution occurs not only between pairs, but also between multiple interacting entities, a concept known as diffuse co-evolution [Pazos and Valencia, 2008; Carmona et al., 2015], that tends to take place within populations. As is the case with evolution, the central idea of co-evolution has a molecular basis, and can be traced to the level of nucleotide mutations [Lynch, 2023], with the main difference being that co-evolutionary changes are brought about by reciprocal selective forces imposed by interaction between two evolving bodies, as opposed to independent selection from potentially shared pressures that can come from external, environmental sources. The bulk of studies on co-evolution focuses internally, on correlated amino acid changes and protein co-evolution between the proteins within a certain species, but relies on comparison to homologous proteins from different species to extract co-evolutionary patterns. Such patterns are identified at the amino acid residue level as changes in one or multiple residues that bring about a change in residues at another site, which can be within the same protein (intra-protein covariation), or between proteins (inter-protein covariation). The more that residue pairs are seen to covary, the higher the likelihood that they are functionally relevant to each other, often at an interaction interface (inter-protein) or to maintain the correct fold of a protein (intra-protein).

The search for evidence of protein co-evolution is not straightforward, since already in the simplest, pairwise case, it requires deducing two *dependent* evolutionary trajectories from proteins suspected to interact. This represents several layers of stacked assumptions, not all of which are easily verifiable. For instance, the base assumption of protein-protein interaction may be demonstrated in one species through experimental means, and yet unknown in related species, so searching for co-evolutionary signals from sequence comparisons between the purportedly interacting protein families often runs into confounding information. For example, in a large-scale case study of potential coevolution in yeast proteins, correlated evolution was found but not due to co-evolution and instead was attributed to shared function within a certain pathway without direct interaction [Hakes et al., 2007]. Additionally, proteins within an organism often do not evolve at the same rates [Zhang and Yang, 2015], therefore co-evolution between two proteins is not necessarily symmetric. Due to these difficulties, true co-evolution is hard to identify, and some experts argue that co-evolution is overrepresented, or not as well demonstrated with empirical evidence as evolution overall [Carmona et al., 2015].

From a practical standpoint, however, whether co-evolution is really behind observed correlated evolution is often not the point; rather, it is the the utility of co-evolution as a conceptual framework to study protein-protein interaction and protein function overall that has revolutionised protein research. This utility was demonstrated early on [Pellegrini et al., 1999] and has been confirmed independently on multiple counts thereafter [Süel et al., 2003; Lovell and Robertson, 2010]. One of the earliest studies identified co-evolution as a correlated presence-absence pattern of proteins across related species and used this correlation to successfully assign function to previously unknown proteins [Pellegrini et al., 1999]. This established that co-evolutionary patterns on the whole-protein level can be detected using phylogenetic techniques. Focusing on correlated amino acid changes in protein domains, another study found such covariation to be indicative of structural proximity and therefore useful to identify binding sites or potential binding partners within a protein complex [Yeang and Haussler, 2007]. Current state-of-the-art protein structure predictors use co-evolutionary information to narrow the search space of potential structures constructed from biophysical neural network models [Roney and Ovchinnikov, 2022]. Altogether, the use of co-evolution as a search for covariation between protein sequences has proven critical in supporting protein evolutionary and functional studies.

1.3 Computational advances in detecting protein co-evolution: from mirrortree to direct coupling analysis

The search for evidence of protein co-evolution is done mostly computationally, as a mining of evolutionary sequence information. This is partly due to the fact that the time scale on which co-evolution takes place is not easily reproduceable in a lab setting, even for rapidly reproducing organisms, limiting experimental studies of co-evolution. Furthermore, the volume of sequence data and the scale on which co-evolution information can be extracted requires computational approaches, which effectively guide experimentation and therefore complement wet-lab co-evolutionary studies. Computational co-evolutionary methods can be divided between phylogenetic methods, global statistical models of sequence alignments, and neural-network based learning of sequence covariation, all of which rely on large databases of protein sequence data to identify co-evolutionary patterns. The timeline of development of these methods thus follows closely the development of high throughput sequencing methods in the last three decades. Generally, the more sequences there are for a taxonomic group, the better the sequence coverage of that group in terms of how well the evolutionary picture of the group can be reconstructed.

The first generation of co-evolutionary methods, called *mirrortree* methods, were developed following a surge of sequence availability in the early 2000s. Mirrortree methods stemmed from observations that proteins suspected to co-evolve had similar evolutionary histories, as seen in the branching pattern of their phylogenetic trees [Pazos and Valencia, 2001; Ramani and Marcotte, 2003]. In other words, the evolutionary distances between the proteins was similar enough that the proteins are suspected to have evolved together. This was attributed to the fact that interacting proteins are bound by similar evolutionary constraints in terms of the pressure to maintain functional interaction, and participation in the same biological pathways thus binds their evolutionary histories together. Through this, the rate of change in interacting partners is influenced from two sides, in the sense that changes to one partner select for and bring about changes to the other. Such methods are essentially a comparison of correlation between phylogenetic distances. Mirrortree methods have been used to successfully identify novel protein-protein interactions [Juan et al., 2008; Zhou and Jakobsson, 2013], and also at the molecular level, to identify interaction sites [Dou et al., 2006]. However, a critical limitation of this method is that the assumption of co-evolution is not always correct. Using similarity of evolutionary trajectories as a proxy for existence of co-evolution is incomplete, as it doesn't directly test for co-evolution at the molecular level, nor does it account for the influence of background evolution of the organism harboring the protein, which tends to bias phylogenetic branching. The case mentioned earlier where independent evolution occurs within a similar environment certainly applies here. Mirrotree methods essentially detect the balance between conservation of sequences and (potential) co-evolution of sequences, and have been found to perform best within a slightly wider span of evolutionary time scale [Zhou and Jakobsson, 2013], limiting its applicability for proteins on a narrower evolutionary time span, that have diverged less from their ancestral interacting forms.

The second approach to identify protein co-evolution, developed in the 2010s, involves searching multiple sequence aligments (MSAs) for covariation patterns, and fitting an approximate statistical model that accounts for the patterns of covariation observed. Columns in the multiple sequence alignment represent positions within the protein, and covariation is detected between two columns as corresponding mutation events, occurring in the homologous proteins of related species (see Fig. 1). This means that if a significant number of amino acid substitutions occurred at some position, and a matching pattern of substitutions is found at another position in the same species, co-evolution could have been responsible. The two terms, *covariation* and *co-evolution*, are not interchangeable - and are used here to refer to the evidence of correlated mutations in the MSA and the overall evolutionary mechanism behind this, respectively. Sequence covariation-based approaches attempt to directly extract co-evolving signals from a sequence-based evolutionary overview of a protein, and were developed first in the single protein case, to study protein folding, and then extended to the interprotein case, to study protein-protein interactions (PPIs). In the first case, covariation serves as an indicator of residues in proximity in the protein, such that these residues maintain the structure of the protein, whereas in the second case, covariation indicates residues important for PPI. The type of protein co-evolution detectable through covariation is between regions that have some variability and are not highly conserved. Many functional regions within and between proteins are highly conserved, and therefore covariation-based co-evolutionary methods are known to show some bias to background conservation levels, and report inconsistent levels of covariation [Fodor and Aldrich, 2004].

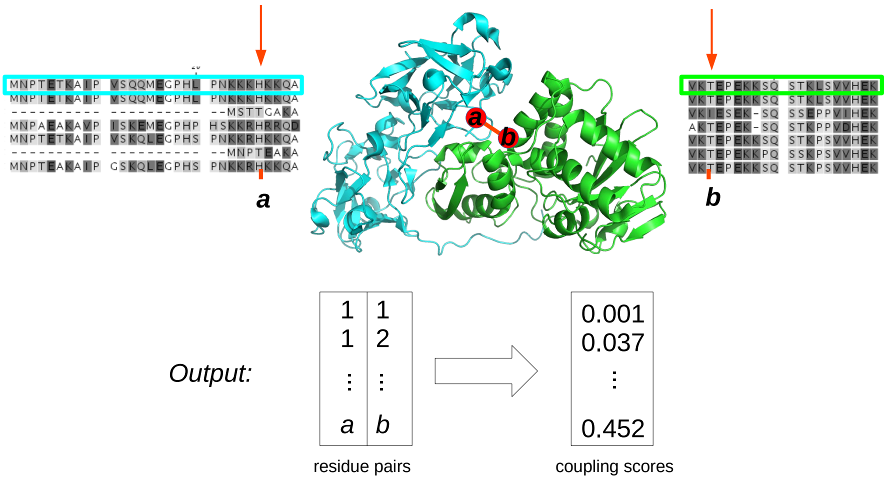

Figure 1: **Illustration of covariation in MSA:** How co-evolutionary methods work: reciprocal mutations, or covariation, detected between positions a and b are given a score according to a global statistical model that calculates coupling scores across all pairwise combinations of residues in this inter-protein case. MSAs corresponding to the proteins are labeled with the target sequence in a colored box, with the color matching the structure. Structures were rendered via PyMOL Schrödinger, LLC [2015].

The earliest sequence-based covariation methodology borrowed the concept of mutual information (MI) from information theory to map covariation within an MSA, by comparing distributions of amino acids in two MSA positions to detect if one position's distribution was dependent on the other [Dunn et al., 2008; Schug et al., 2009; de Juan et al., 2013]. This was initially done irrespective of the types of amino acids in the distributions, and therefore did not incorporate information on biologically relevant amino acid substitutions [Dunn et al., 2008]. Covariation detected by MI was shown to generate many false positives, as it was easily impacted by random variations or noise, phylogenetic background biasing residue distribution, and low sequence availability [Martin et al., 2005; Weigt et al., 2009]. One important shortcoming that MI methods brought to attention was that covariation could arise between residues that were not directly interacting (e.g. allosteric regulation), but indirectly covarying due to sharing interactions with a third residue, or even to multiple other residues. MI methods detect covariation on a pairwise level, and were lacking a global overview of covariation patterns across the entire protein. The next generation of co-evolutionary methods focused on solving this issue, and did so first with global inference based on the message-passing algorithm from statistical physics [Weigt et al., 2009], and this methodology was eventually named *direct coupling analysis*.

Direct coupling analysis (DCA) expands on MI methods by introducing a residue de-coupling step based on the generalised Potts model, which essentially assigns a probability to each sequence in an MSA based on detected 'spins' of 20 possible orientations (20 standard amino acids), calculated over all pairwise combinations of positions in the MSA. The parameters of this model make a statement about first the tendency of an amino acid to be found at a certain position in the MSA, and then gives a numeric value associated with the tendency of a position to be dependent on another position, for every other position possible in the MSA. The model is therefore fully connected, and each pair is assigned a *coupling score*, which represents the strength of covariation between those two positions, as seen relative to the entire scope of covariation present in the MSA. One key bottleneck in the use of DCA was that using the generalised Potts model requires the calculation of a normalisation constant (to ensure all sequence probabilities summed to 1), which, for large MSAs, becomes computationally intractable (see Section 2). The first formulation of DCA [Weigt et al., 2009] approximated this normalisation constant using the message-passing algorithm, and further development of DCA focused on optimising this approximation process to make DCA applicable to larger proteins and deeper MSAs. This led to many flavours of DCA: mpDCA [Weigt et al., 2009], mfDCA [Morcos et al., 2011], gaussDCA [Baldassi et al., 2014], and plmDCA [Ekeberg et al., 2013]. What is notable about DCA as a co-evolutionary method is that, compared to previous methods, it performed consistently well in benchmark studies on bacterial proteins or protein complexes, and was able to accurately identify residue-residue interactions that were important to maintain protein folding or PPI [Uguzzoni et al., 2017]. DCA came to dominate the field, and is still one of the main co-evolutionary methodologies, although its current applications have shifted away from protein structure prediction and protein-protein interactions.

1.4 Advancements in protein structure prediction: the rise of neural network methods

The development of co-evolutionary methods has been a key component of a larger scientific effort to predict protein structure and protein-protein interaction at the residue-level, and while mirrotree, MI, and DCA methods contributed a significant amount to this effort, the recent application of deep learning methods, in particular of neural network (NN) architectures, outperformed all previous methods and completely reset the field. Successful NN-based methods have taken advantage of the growth in protein sequence data and the improvement of computational hardware and resources to train the predictors to not only learn co-evolutionary information but also biochemical constraints and other features important in the protein structural and interaction prediction problem. NN-based methods are therefore much more effective at extracting information from an MSA, compared to previous co-evolutionary methods. In fact, NN-methods constitute a category of their own, and have developed beyond co-evolutionary prediction. One NN-method relevant to this thesis is Google Deepmind's AlphaFold2 [Jumper et al., 2021], and it will be explored more in depth in Section 2 and Paper 3.

1.5 Test case 1: NAD pathway from protein perspective

Nicotinamide adenine dinucleotide (NAD) and its phosphorylated version NADP are essential coenzymes for metabolism and many other bioenergetic processes in all living organisms. NAD serves a catabolic function, mainly as a coenzyme in redox reactions, cycling between oxidised NAD$^+$ and reduced NADH forms to ultimately synthesise and transfer energy in the form of ATP to fuel other pathways [Chiarugi et al., 2012]. NADP also shuttles electrons but mainly serves an anabolic function, also to maintain cellular redox homeostasis.

From its redox role, NAD is involved in glycolysis, fatty acid oxidation, and the citric acid cycle [Ziegler and Niere, 2004]. NAD also plays a role in intracellular signal transduction pathways, influencing DNA repair and gene expression, and when this is the case it is consumed and converted to nicotinamide (Nam). To fuel the multitudes of NAD-consuming reactions, cells need to maintain sufficient pools of NAD, and therefore intracellular NAD concentration is tightly regulated [Ruggieri et al., 2015]. Replenishing of NAD levels is accomplished through NAD *de novo* synthesis and NAD salvage from Nam. Cellular strategies to maintain NAD levels vary throughout the taxonomic kingdoms and also cell and tissue types [Ding et al., 2021; Amjad et al., 2021], but generally consist of a mix of biosynthetic and salvage pathways, such that some level of redundancy is preserved.

Two biosynthesis pathways have been identified for NAD. The Preiss-Handler pathway starts from nicotinic acid (NA), obtained through dietary means from niacin, which is phosphoribosylated to nicotinic acid mononucleotide (NAMN) by nicotinate phosphoribosyltransferase (NAPRT), followed by NAMN adenylation to NAAD by NAMN transferase (NMNAT) and finally NAAD amidation NAD$^+$ by NAD synthase [Preiss and Handler, 1958; Ruggieri et al., 2015]. The Preiss-Handler pathway is by far the most studied NAD biosynthetic pathway, and is present in both prokaryotes and eukaryotes [Bockwoldt et al., 2019]. The second biosynthetic pathway, the *de novo* synthesis pathway, starts from either of the amino acids: aspartate (plants and bacteria), tryptophan (mammals, fungi, other bacteria), and converges at quinolinic acid (QA) to undergo a three-step conversion to NAD, which represents the NAMN-NAAD-NAD steps of the Preiss-Handler pathway.

In addition to these pathways, NAD salvage from Nam is an important source of NAD recycling back into the system, and constitues most of the newly synthesized NAD$^+$ in human cells [Chiarugi et al., 2012]. Two salvage pathways have been identified: a two-step pathway involving the phosphoribosylation of Nam into Nam adenine dinucleotide (NMN) by nicotinamide phosphoribosyltransferase (NAMPT), and the adenylylation of NMN to NAD by NMNAT, and a four-step pathway that starts with conversion of Nam to NA by nicotinamidase (PncA), followed by the three reactions of the Preiss-Handler pathway. The two-step salvage pathway is found to dominate in vertebrates, and is characterised by a high affinity of NAMPT towards Nam, such that nearly all of the Nam present is recycled in to NAD. This makes it more energetically efficient than the four-step pathway. Previous work on the evolutionary trajectories of these two pathways [Gazzaniga et al., 2009; Gossmann et al., 2012; Bockwoldt et al., 2019] showed a mixed picture, with preference for the four-step pathway in many bacteria, but presence of the two-step pathway in other prokaryotes and sometimes coexistence of both pathways in prokaryotes and eukaryotes.

Disregulation of NAD metabolic and signaling pathways is implicated in tumor formation and cancer phenotypes of various tissues, as well as numerous age-related disorders, and a significant amount of biomedical research focuses on targeting enzymes that have regulatory roles on intracellular NAD levels. Due to the importance of NAD salvage from Nam in mammals as a way to maintain constant NAD supply, the salvage pathway is a prime target for such therapies. Metabolic modeling efforts, combined with experimental evidence, have identified key steps of the two-step pathway and the biosynthetic pathway as regulatory targets [Ruggieri et al., 2015]. Gaining an evolutionary perspective on the various strategies employed by organisms to maintain NAD homeostasis is crucial in further advancing the NAD pathway as a cancer treatment target.

1.6 Test case 2: mTOR signaling from a protein perspective

The eukaryotic mechanistic target of rapamycin (mTOR) signaling pathway serves as a central regulatory hub of many anabolic cellular processes such as cell growth and proliferation, and catabolic processes such as autophagy and apoptosis. The mTOR pathway translates both internal and external factors in the form of growth factors, amino acids, energy status, and oxygen levels, among other factors, to control gene transcription, protein and lipid synthesis, mitochondrial metabolism and biogenesis, and many other processes related to cell growth [Laplante and Sabatini, 2009; Zou et al., 2020]. Naturally, then, mTOR pathway deregulation and general dysfunction has been implicated in tumor growth and cancer, and much of current research focuses on mapping out the entire mTOR pathway and identifying the functions of subpathways as targets for cancer-fighting therapeutic purposes [Saxton and Sabatini, 2017].

The mTOR signaling pathway is large and operates in a multi-compartmental arena, consisting of protein-protein interactions of 241 unique proteins, and protein interactions with genes, RNA, and numerous other molecules across intracellular and extracellular zones [Caron et al., 2010]. The entirety of mTOR pathway activity is organised around two protein complexes, mTOR complex 1 (mTORC1) and mTOR complex 2 (mTORC2). These complexes are both named after their central component, the mTOR protein, a serine/threonine kinase from the PI3K-related kinase family, which forms the catalytic subunit of these two complexes. mTORC1 and mTORC2 differ in their protein composition of regulatory partners and general function - mTORC1 promotes cell growth through lipid and nucleotide synthesis, protein synthesis, and control of autophagy, and mTORC2 functions as an insulin-dependent regulator of cell survival and proliferation [Saxton and Sabatini, 2017]. The functions of mTORC1 have, compared to mTORC2, been more comprehensively determined. However, a full characterisation of the entire pathway, with computational modeling and experimental validation, is yet to be completed.

This thesis focuses on selected protein components of three subpathways in mTOR signaling involving mTORC1 function and regulation, from the perspective of interaction site and structural modeling. The first process involves protein synthesis via mTORC1 phosphorylation of the eukaryotic initiation factor 4E (eIF4E)-binding proteins (4EBP1, 4EBP2). Phosphorylation of the 4EBPs changes their mode of binding to eIF4E, allowing for mRNA translation to occur [Böhm et al., 2021]. The proteins examined here are constituents of the 4EBP:eIF4E complex. The second process involves the tuberous sclerosis (TSC) complex, an important negative regulator of mTORC1 and therefore a control switch on all downstream cell growth-promoting processes [Huang and Manning, 2008]. TSC complex-mediated regulation of mTORC1 occurs at the lysosome [Prentzell et al., 2021; Rehbein et al., 2021], through the binding of a component protein of the TSC complex, TSC2, with a tethering protein, the Ras GTPase-activating protein-binding protein 1 (G3BP1). From this subpathway, we therefore focus on G3BP1:TSC2 binding. The final protein complex examined involves a subcomponent of mTORC1, Raptor, and its regulatory substrates 4EBP1 and S6 kinase 1 (S6K), that are responsible for nutrient sensing and also mTORC1 inhibition [Kim et al., 2002]. These protein complexes are examined for their regulatory relevance in mTOR signaling and for their potential co-evolutionary history, as well as for the challenging protein characteristics when it comes to structural modeling.

1.7 Thesis objectives

Protein evolution and co-evolution are two key theories that form the basis for and continue to shape research into protein function at the molecular, cellular, and whole-organismal levels. The primary objective of this thesis is to build upon existing bioinformatic methodology for identifying protein co-evolutionary patterns, both on the taxonomic, species-level scale, and also at the molecular level, in PPI. We thus first develop a novel tool to phylogenetically map protein presence-absence in the context of studying metabolic and signaling pathways: a map that can be used as a large-scale overview of evolutionary and potential co-evolutionary patterns. Then, we develop a pipeline to incorporate DCA and detect co-evolution between dimerising proteins, with the aim to pinpoint potential interacting regions in the PPI context. We also construct a novel visualisation scheme for co-evolutionary results. Finally, we benchmark AlphaFold2, the current state-of-the-art in protein structure prediction, which has been trained on co-evolution information, through testing its ability to handle prediction of PPI interfaces for the difficult use case of a large and dynamic protein complex.

More specifically, the aims of this thesis are, as outlined by research question and the corresponding objective(s) per paper:

How can co-evolution and evolution of proteins be mined out of the current landscape of protein sequence data in such a way that association patterns can be compared across many organisms?

Objective 1 (Paper 1): Develop and present a tool to extract protein presence-absence information from sequence databases to span the entirety of the taxonomic tree and output results that enable users to identify large-scale patterns for groups of proteins, for the purpose of studying taxa-specific biological pathway preferences arising from evolution or co-evolution. Test tool on a specific biological use case.

Objective 2 (Paper 2): Apply the tool from Paper 1 to another biological test case, to explore if there are temperature-dependent NAD recycling pathway preferences across extremophilic bacterial clades, or habitat-dependent preferences in bacterial clades associated with eukaryotic hosts.

For detecting protein sequence co-evolution, how viable is DCA as a means of identifying inter-protein co-evolution for the purpose of finding PPI at the residue-level?

Objective 3 (unpublished results): Develop a tool that wraps DCA into a pipeline for dimeric co-evolutionary detection, and test if detected co-evolution is a reliable indicator of interaction in protein pairs from the mTOR pathway. Create a set of visualizations to represent co-evolutionary scores focusing on comparison of score amplitudes and residue pair connectivity.

Given that protein co-evolution information has been incorporated into the training of AlphaFold, *can the algorithm reliably predict protein-protein interaction for non-standard test cases?*

Objective 4 (Paper 3): Develop strategies to use AlphaFold Multimer to predict protein-protein interaction interfaces on disordered and dynamic protein complexes from the mTOR signaling pathway, to test if the prediction algorithm can identify interactions at the residue-level.

2 Materials and methods

The methodology for this thesis work is divided into three categories: large-scale extraction of protein presence-absence information, protein sequence co-evolutionary detection, and protein structure and PPI prediction. These methods are computational, and drew data from publicly-available databases from the National Center for Biotechnology Information (NCBI) [Dat, 2016] and its constituent sequence database hubs. These methods were also inspired from collaborations with biological laboratories in Bergen (Ziegler lab) and Innsbruck (Thedieck lab). Section 2.1 discusses these partnerships, and the origin and use of data in this thesis work overall. Sections 2.2 - 2.4 give technical outlines of the three methodological categories: phylogenetic presence-absence analysis, co-evolutionary pipeline with direct coupling analysis (DCA), and structural prediction with AlphaFold [Jumper et al., 2021]. Bioinformatic work, especially that which involves large-scale sequence queries and neural network constructions, is reliant on computing systems to support analysis. Section 2.5 thus presents the computational setup that made the analysis in this work possible.

2.1 Data

Protein sequence data and structural data were extracted from online databases hosted by the NCBI, with emphasis on ensuring correctness and quality of the data by using human-verified annotated sequences whenever possible. Target protein sequence data was taken from the SwissProt UniProtKB database [Apweiler et al., 2004] as verified protein sequences, and large-scale queries of up to 5000 related sequences were submitted against the NCBI non-redundant (*nr*) database. These sequences span the entire taxonomic tree of life. To verify predicted contacts and/or structure, reference protein structure and sequence data were taken from the Protein Data Bank [Berman et al., 2000]. Through these online sources, the majority of data used in this project is publicly available and made accessible with the versions of databases used in queries specified in the respective manuscripts.

Collaborations with Mathias Ziegler's lab at University of Bergen, and Kathrin Thedieck's lab at the University of Innsbruck provided data in the form of experimentally-driven inspiration for this project, particularly in guiding the choice of biological pathway and protein targets to explore with co-evolutionary analysis and for phylogenetic analysis and modeling with AlphaFold. The Ziegler lab focuses on the NAD pathway and cellular bioenergetics, and therefore influenced the use case of NAMPT and PncA in Paper 2, whereas the Thedieck lab focuses on the mTOR pathway and cellular metabolism, and thus provided support and ideas for the modeling of the 4EBP complex and the TSC complex in Paper 3, and the Raptor complex in our unpublished results.

2.2 Phylogenetic presence-absence analysis

The methodology for the large-scale extraction of protein presence-absence information is comprehensively described in Paper 1 as a software tool, VisProPhyl, therefore this section only presents the workflow diagram for the pipeline (Figure 2), followed by a brief overview of the methods, and the reader is referred to Paper 1 for more detailed descriptions of each step.

The VisProPhyl tool is available on GitHub at `https://github.com/MolecularBioinformatics/VisProPhyl`. The main workflow is written into the *phylogenetics.py* script, which is as follows: Target protein sequences of interest (usually human sequences) are taken from the functionally verified UniProt Swiss-Prot database (release 2022.4), and submitted as queries via Blastp [Altschul et al., 1990; Camacho et al., 2009] to the NCBI nonredundant (nr) protein database (release 2022.2.21) [Dat, 2016]. Blastp parameters are kept at defaults (BLOSUM62 matrix, word-size 6, gap-open cost 11, extension cost 1, and e-value threshold 0.001) with the exception of an increase in maximum number of hits returned to 5000. Sequence sets consisting of the target protein with its blasted hits are filtered for false positives based on implementing a minimum length cutoff per protein and an e-value cutoff (1e-30) based on cross-hits. These sequence sets are taken to be representative of a protein's pattern of spread over the taxonomic groups. Using ETE Toolkit [Huerta-Cepas et al., 2016] and NCBI's taxonomic tree [Schoch et al., 2020], the blast results are combined and mapped to a phylogenetic tree, and also to a heatmap for a zoomed-in view on presence / absence patterns for model organisms. Scripts for the analysis are written in Python 3.5.

2.3 Protein sequence co-evolutionary workflow

We developed a pipeline to perform protein sequence co-evolutionary detection for this thesis work, tying together methodology for sequence querying, multiple sequence alignment, and direct coupling analysis, with the necessary input and output (IO) processing steps in between. This pipeline detects dimeric, 2-body inter-protein coevolution, as well as any intra-protein coevolution individually in either of the partners. Code for this pipeline can be found at: `https://github.com/MolecularBioinformatics/euk_dim_dca.git`. The procedural steps for this pipeline (corresponding to the script *run_workflow.py* in the GitHub repository) are given as a workflow diagram in Fig. 3, together with columns indicating file output formats and technical descriptions of each step.

As the co-evolutionary methodology is not factoring into Papers 1-3, it is described in detail here. The co-evolutionary workflow consists of 10 steps and is divided into 5 intermediate stages (grouped by color in Fig. 3) : (1 - gray) collection of target protein reference sequences, (2 - brown) homolog search through phmmer [Finn et al., 2011], (3 - red) sequence extraction and quality check, (4 - blue) alignment creation with Muscle [Edgar, 2004] and post-processing, and (5 - violet) computation of mfDCA coupling scores with pydca [Zerihun et al., 2020]. Following the pipeline is a post-processing and visualisation scheme.

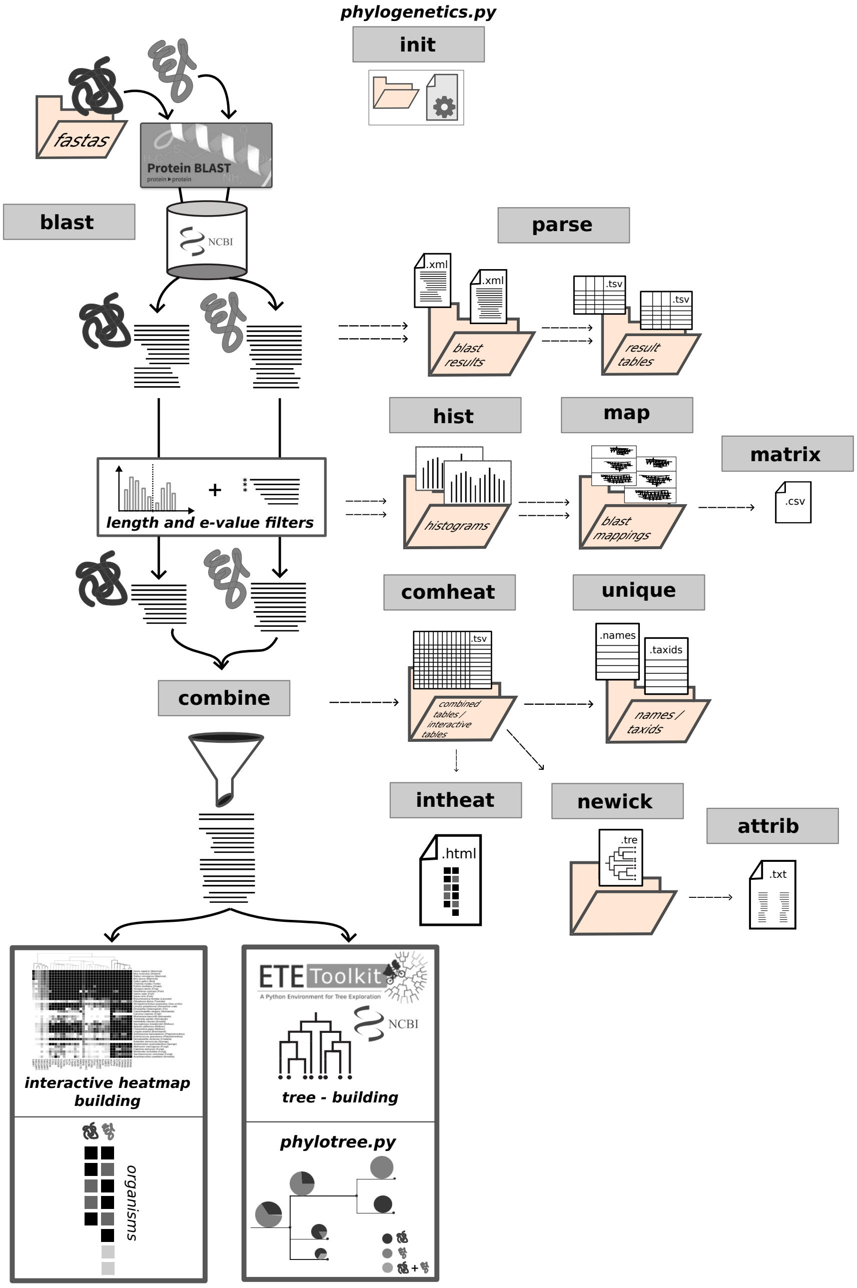

Figure 2: **Workflow Diagram for Phylogenetic Presence-Absence Analysis:** All steps correspond to the *phylogenetics.py* script in the GitHub repository: `https://github.com/MolecularBioinformatics/VisProPhyl`. The left-most column shows the main procedure of sequence file querying, sequence hit filtering, and result combination, followed by heatmap and tree visualisations. Arrows to the right show the file outputs per step, and their location in the folder scheme. Refer to Paper 1 for an in-depth methodological explanation of the workflow.

2.3.1 Stages 1 - 3

Accurate detection of sequence coevolution depends first on finding potentially homologous sequences related to target proteins. The search for homologous sequences was carried out by

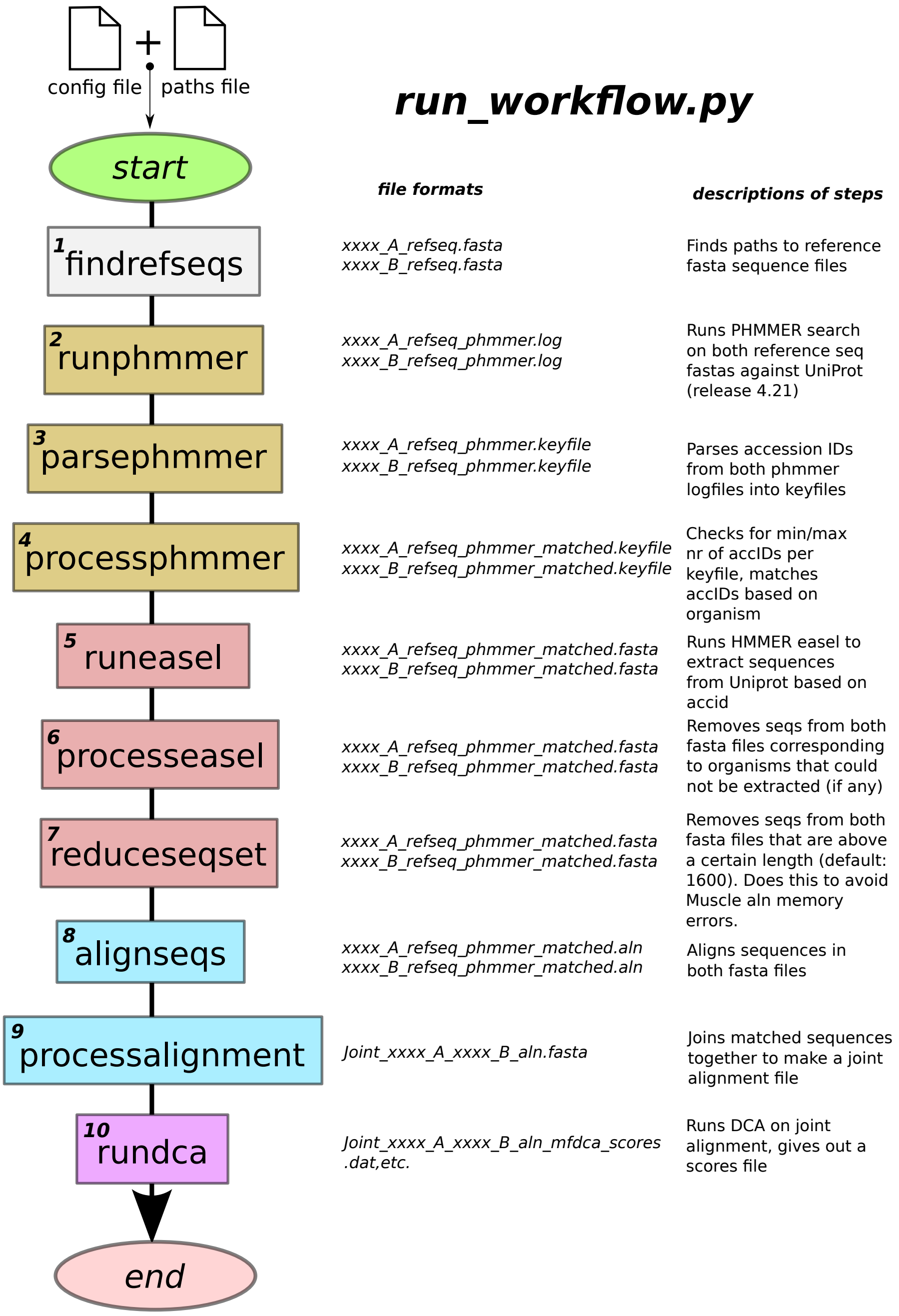
config file
paths file
run_workflow.py
start
file formats
descriptions of steps
1 findrefseqs
xxxx_A_refseq.fasta
xxxx_B_refseq.fasta
Finds paths to reference fasta sequence files
2 runphmmer
xxxx_A_refseq_phmmer.log
xxxx_B_refseq_phmmer.log
Runs PHMMER search on both reference seq fastas against UniProt (release 4.21)
3 parsephmmer
xxxx_A_refseq_phmmer.keyfile
xxxx_B_refseq_phmmer.keyfile
Parses accession IDs from both phmmer logfiles into keyfiles
4 processphmmer
xxxx_A_refseq_phmmer_matched.keyfile
xxxx_B_refseq_phmmer_matched.keyfile
Checks for min/max nr of accIDs per keyfile, matches accIDs based on organism
5 runeasel
xxxx_A_refseq_phmmer_matched.fasta
xxxx_B_refseq_phmmer_matched.fasta
Runs HMMER easel to extract sequences from Uniprot based on accid
6 processeasel
xxxx_A_refseq_phmmer_matched.fasta
xxxx_B_refseq_phmmer_matched.fasta
Removes seqs from both fasta files corresponding to organisms that could not be extracted (if any)
7 reduceseqset
xxxx_A_refseq_phmmer_matched.fasta
xxxx_B_refseq_phmmer_matched.fasta
Removes seqs from both fasta files that are above a certain length (default: 1600). Does this to avoid Muscle aln memory errors.
8 alignseqs
xxxx_A_refseq_phmmer_matched.aln
xxxx_B_refseq_phmmer_matched.aln
Aligns sequences in both fasta files
9 processalignment
Joint_xxxx_A_xxxx_B_aln.fasta
Joins matched sequences together to make a joint alignment file
10 rundca
Joint_xxxx_A_xxxx_B_aln_mfdca_scores .dat,etc.
Runs DCA on joint alignment, gives out a scores file
end

Figure 3: **Workflow Diagram for Co-evolutionary Analysis:** All steps correspond to the *run_workflow.py* script in the GitHub repository: `https://github.com/MolecularBioinformatics/euk_dim_dca.git`. The left-most column shows steps from preliminary input file preperation through the start to the end, passing through five colored stages for a total of 10 steps in total. The general flow follows target sequence identification, sequence querying to identify homologs, extraction of reference sequences for the homologs, filtering of homolog sets based on common organisms across the two sets, multiple sequence alignment of homologs to the target, and co-evolutionary analysis. The middle column details output file formats, and the right-most column gives a concise technical description of each step.

submitting target protein sequences as queries with the phmmer [Finn et al., 2011] database search tool, which performs database search with Hidden-Markov model profiles, and queried against the UniProtKB/SwissProt database (release v. 2019_11) [Apweiler et al., 2004]. The parameters used in the phmmer runs were left at default (i.e. gap open probability 0.02, gap extend probability 0.4, substitution matrix BLOSUM62), the inclusion threshold (E-value $\leq$ 0.001), and the maximum number of iterations set at 5. To obtain a preliminary set of putative homologs, the first 5000 hits were taken for each of the target protein queries. A basic quality check on E-value ranges and bit scores was performed on these hits, and sequences scoring over the E-value threshold were discarded. The full sequence bit score and bias correction values were compared per sequence to make sure high-scoring non-homologs were not retained. Using the accession IDs of these two sets of top 5000 hits, the corresponding full sequences were extracted from the UniProt database using Hmmer's easel miniapp toolkit.

An important step after extracting and quality checking homologous sequences is to determine if the homologous sequences from either protein partner are themselves interacting, such that each line of the multiple sequence alignment created from the queried set contains a pair of interacting proteins. Ensuring interaction is currently an open problem, as the verification of interaction lags behind sequence discovery, and is especially tough for eukaryotic sequences that have no operon-like genomic location to use as proxy for interaction. At the very least, paired sequences should originate from the same organism. In the pipeline, the sets of 5000 top-scoring sequences are thus cross-examined for common organisms. One representative sequence was retained per organism (per protein), which was chosen to be the longest sequence, to avoid selecting subsequences. Then, the effective number of sequences (Neff) is calculated, and average sequence identity is checked. At this point, two equally sized sets of sequences have been curated, with homologs to each target sequence that originate from the same set of organisms. These two sequence sets are aligned and then paired by organism in the next stage.

2.3.2 Stages 4 - 5

Muscle (v3.8.21) was used to build an MSA per protein 'family', consisting of the target protein, aligned with its potential homologs. Resultant alignments were then trimmed using the target protein as a reference, to cut down on the number of gaps in the alignment (pydca trim_by_refseq). At this stage, both alignments contain paired sequences of the corrected length, and are submitted to pydca for calculation of co-evolutionary values.

Of the various direct coupling analysis (DCA) frameworks available today, mfDCA is preferred for its ability to handle computationally-intensive, high-throughput tasks, with minimum sacrifices in terms of predictive accuracy. The implementation of mfDCA from pydca Zerihun et al. [2020], a Python-based DCA suite, is used here to calculate residue-residue co-evolutionary scores. Co-evolutionary scores generated through this pipeline are further analysed by taking slices of the top 100 - top 500 highest scoring residue pairs, and visualised on the structure (if a structure is available) or as parallel plots or hairball plots, which are demonstrated in Section 3.

2.3.3 Post-processing: Visualisation of co-evolutionary scores and MSA analysis

This co-evolutionary pipeline returns DCA scores for every pairwise combination of residues, therefore only the top scoring residue pairs are taken for further analysis and visualisation. By convention, the top 100 to top 500 scoring residues are analysed. DCA score ranges are not intercomparable between separate co-evolutionary analyses, meaning that the strengths of co-evolutionary signal between a pair of residues is relative only to other residue pairs within the same protein(s) of the co-evolutionary analysis. Furthermore, mfDCA as implemented in [Zerihun et al., 2020] does not normalise scores, therefore score ranges can vary from -1 to +3 in our analyses. It is not the absolute value but the relative value of the score that is taken for further interpretation.

Then, to complement DCA scores, we compute conservation scores per column of the MSA. High conservation typically inhibits substantial co-evolution, whereas high variability fosters it. Hence, conservation scores act as a validation measure for the DCA methodology when applied to individual residue pairs. ConSurf [Ashkenazy et al., 2016], a bioinformatic tool that infers functional domains from nucleic acid or amino acid sequences, is used to compute the conservation scores. ConSurf ties together sequence homology search via BLAST or CS-BLAST2 [Altschul et al., 1990], multiple sequence alignment with MAFFT [Katoh et al., 2002], and phylogenetic tree building via Rate4Site4 [Mayrose et al., 2004] to compute the evolutionary rates and conservation scores used in its functional domain inference. ConSurf outputs the original query sequence, with each residue ranked 1-9 from low to high conservation, and also labeled with predicted solvent accessibility (exposed/buried) and also its inferred purpose (functional/structural). The ConSurf web server `https://consurf.tau.ac.il/consurf_index.php` was used in this analysis.

As a summary of results, DCA score amplitude, residue pairing, and conservation, are visualised using a novel series of plots (see Fig. 4). Hairball network visualisations (leftmost panel, Fig. 4) show linkage patterns of interprotein residue pairs scoring within a defined range of co-evolutionary values (e.g. top 100, 300). In a hairball plot, nodes represent amino acids, numbered according to their sequence positions, and coloured based on the protein of origin, and edges represent connections, or the links between amino acids that may have undergone co-evolution. Node sizes reflect the degree, or the number of outgoing connections made by that amino acid to residues from the other protein. The plots are created as bipartite graphs from the R package *igraph* [Csárdi and Nepusz, 2006], with visualisation layout 'fruchterman reingold'. Score histograms are created with MatPlotlib [Hunter, 2007] to show the distribution of DCA scores, and also to indicate ConSurf conservation scores together with parallel plots (rightmost panel, Fig. 4) from Plotly [Inc., 2015], which show highly co-evolving residues between two proteins, where the indices for each protein's residues are displayed on either vertical axis, and lines are drawn where co-evolution is detected, with the coloring of line indicating the magnitude of the score.

The final post-processing step is to extract information from the input MSAs, to better understand sequence depth and taxonomic composition of our input. Calculations of the effective number of sequences (equivalent to the number of sequence clusters at 62% sequence similarity) are compared to the total number of sequences. Then, metadata is extracted from UniProt with information on the species of origin of each sequence, to search for potential taxonomic bias in the MSA.

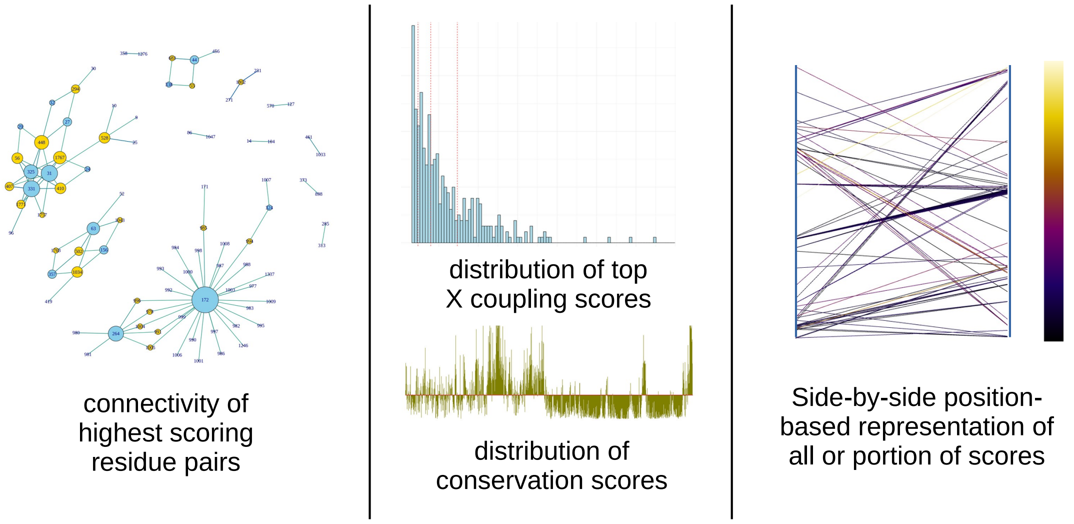

Figure 4: **Visualisation Plots of Co-evolutionary Scores:** Hairball network visualizations are utilized to depict linkage patterns among interprotein residue pairs that score within a specified range of co-evolutionary values. Nodes are amino acids, node sizes represent the degree of each node, and nodes are colored based on the protein they originate from. These visualizations are generated as bipartite graphs. Histograms illustrate the distribution of DCA scores and conservation. Additionally, parallel plots highlight highly co-evolving residues between two proteins, with each protein's residue indices displayed on separate vertical axes. Lines indicate detected co-evolution, with line color indicating the magnitude of the score.

2.4 Protein structure and PPI prediction

Prediction of protein structure and PPI interfaces was done with Google DeepMind's AlphaFold2 (AF2) [Jumper et al., 2021] to predict single protein structures, and AlphaFold Multimer (AF2M) [Evans et al., 2022], to predict structures of protein complexes (and PPI). This work used AF2 and AF2M version 2.3.1. The AF2 algorithm predicts the structure of a protein based on its amino acid sequence, in the context of evolutionarily related sequences in the MSA. AF2 thus leverages evolutionary and co-evolutionary information from an MSA within an EvoFormer neural network module with learned biophysical constraints in a structural neural network module, and achieves accurate structural predictions through iterative refinement of the entire structure generation process. More specifically, AF2 generates three representations of a predicted structure, a *pair representation* involving template-based homology modeling on the input sequence, an *MSA representation* through the EvoFormer step that extracts correlated mutations and conservation information to construct an MSA-based structural representation, and finally a *structural representation* that ties together the pair and MSA representations to generate a 3D structure. AF2M is an extension of AF2 for predicting structures of multi-protein complexes.

We evaluated AF2 and AF2M predicted structures by comparing them to reference PDB structures for target proteins or protein complexes, whenever such references existed. Comparison of structures consisted of structural alignment via PyMOL [Schrödinger, LLC, 2015] to evaluate similarities between the global folds of references and specific regions of contact. The metric used here is RMSD. Evaluation of structure predictions and structural regions was also done with AF2's own confidence metric, pLDDT.

For our use of AF2 and AF2M in this thesis work, we did not develop a wrapper software as we did in the previous two cases. Code for the post-processing will be made available on GitHub. Paper 3 gives a more extensive description of the methodology for this analysis.

2.5 Computing platforms and HPC job submissions

Computations for this work were carried out over two computing platforms, a local PC and the Saga high-performance supercomputer (HPC) provided by the Norwegian research infrastructure services (NRIS). To give an idea of the computational resources necessary for this project, the specifications of each computer are first given, and then the setup and resource usage for a typical analysis run is described.

The local PC was an HP Probook, running Ubuntu 22.04.4 LTS on 8 Intel Core i5-8265U 1.60 GHz CPU cores and 8.0 GB memory. The Saga supercomputer runs Rocky Linux 9.1 (Blue Onyx) and supports *normal* (non-GPU) computational runs per node on 40 Intel Xeon-Gold 6138 2.0 GHz / 6230R 2.1 GHz CPU cores with 192GB memory per core. Saga also supports *accel* (GPU) computations with 2 Intel Xeon-Gold 6126 2.6 GHz CPU cores with 384GB memory per core and 8 NVIDIA P100 GPU cores with 16 GB per core.

The majority of computational work was run on Saga CPUs (as *normal* jobs), and whenever possible, in batch as parallel runs. GPU usage (*accel* job) was limited to the model-construction step inbuilt into AlphaFold. Due to the procedural nature of much of these analysis methods, most of it was run in a serial manner. Parallelisation of portions of each workflow was achieved through multithreading of sequence querying steps. The following job submissions represent a reasonable upper bound of resource requests tailored to the analysis, such that jobs are not cancelled due to time or memory errors. Many jobs finished well within the bounds of the resources allocated.

An average run of the co-evolutionary workflow, for a protein or protein complex of between 200-1400 amino acids requires 5-10 hours on one node with 4 CPU cores and maximum 25GB of memory, with the sequence query, alignment, and DCA steps taking up around 80% of computational resources.

For the phylogenetic presence-absence analysis, only the sequence query step was run on Saga, as it represented the bulk of the computations for this method. A standard sequence query for a single protein in this analysis runs on one node, 16 CPU cores, for maximum 3 hours and 100MB/core. The remaining tree and heatmap construction steps run locally on one core and for no more than a few minutes.

The protein structure prediction workflow consisted of AlphaFold followed by structural visualisation. Visualisation is done on the local PC, whereas AlphaFold requires intensive computing and ran on Saga, and jobs were divided into single protein (monomer) or multiprotein (multimer) jobs, and further divided based on size (more memory or time). A regular monomeric run is suitable for single proteins up to 700 amino acids, and runs on 1 node, 2 CPU, 1GPU, and 20GB/core, for 8-10 hours. Multimeric runs ranged from: small jobs for total amino acid counts under 500 aa (5 days on 1 node, 4 CPU, 1GPU, 20GB/core) to regular jobs under 1000 aa (10 days on 1 node, 4 CPU, 1GPU, 30GB/core). The largest multimeric jobs for protein complexes of 1000 aa or more required a total of 120GB (14 days on 1 node, 8 CPU, 1GPU, 15GB/core).

4 Discussion

The overall goal of this thesis was to build upon current bioinformatics methodologies for detecting protein evolutionary and co-evolutionary patterns, spanning across the higher taxonomic levels down to species-level analysis, and extending to molecular-level analysis of PPI. From a bioinformatic perspective, building upon previous work involves the development of novel software to streamline workflows by integrating multiple tools and analyses into a cohesive framework. This is primarily done to increase accessibility of the analysis to users within and outside the immediate research field, and in our case, particularly to molecular biologists. Furthermore, from a broader biological perspective, such tools require rigorous testing and evaluation beyond their intended use cases to potential corner cases that may warrant further exploration.

In line with these perspectives, this thesis comprises three studies and additional work that integrate co-evolutionary and evolutionary theories within a computational framework. Altogether, they address two major goals in molecular biology: first, the detection and utilization of enzyme presence-absence patterns to infer biological pathway diversity and organism-specific pathway preferences to reconstruct pathway evolution, or to study functional coevolution of proteins; and second, the prediction of PPI and protein structure across proteins of varying levels of disorder, size, and complexity. The protein targets selected within this framework emerge from biological pathways that have already been extensively studied, laying the groundwork for various unanswered, open questions regarding their specific interactions and functions. Expert knowledge from partnerships with the Thedieck and Ziegler labs established the relevance and steered the focus of these questions, which, together with the computational results from our framework, form the setup for further experimental validation.

In this discussion, we present the precise formulation of the framework that binds together the entirety of this thesis. We then reassess the NAD and mTOR cases using this framework, concentrating on patterns identified at different stages and the resulting insights crucial for subsequent analysis. Following this, we discuss the considerations surrounding protein sequence analysis particular to eukaryotic genetics, then explore result interpretability and how this shapes further validation.

4.1 Adopting a dual-axis framework: extracting co-evolutionary and evolutionary patterns from protein sequence data

The strategies formulated in this thesis constitute components of a larger framework to mine protein evolutionary and co-evolutionary information from sequence data. This framework operates along two axes, combining methodology that differentiates between: explicit or implicit calculations of evolution or co-evolution (axis 1), either on a residue or whole-protein level (axis 2). The idea behind this is that while individual framework components may offer partial insights, their combination more thoroughly probes the question of protein co-evolution and evolution as a whole. VisProPhyl (Papers 1 and 2) represents an explicit, protein-level methodology. VisProPhyl extracts protein presence-absence data from current databases, offering a comprehensive overview of evolutionary forces affecting the protein across many organisms. Similarly, our co-evolutionary pipeline (DCA, unpublished results) is an explicit, residue-level methodology, in that it identifies correlated mutations and overall evolutionary change between residues of protein pairs. Finally, our strategies for use of AlphaFold2 (Paper 3), represent implicit, residue-level analyses, as their use does not involve a direct calculation of co-evolution. Instead, AF2 has integrated co-evolutionary information in its training process, and continues to calculate correlated mutations to predict residue-level interactions and to refine the overall structural prediction.

Organised in this way, this framework tackles progressively narrower biological inquiries related to proteins of interest, where findings from individual inquiries are interconnected and mutually inform one other. This framework begins by determining if one or more proteins have undergone evolutionary change and in which species this change is apparent. A taxonomic map of joint appearance or disappearance of proteins is created at this stage. Then, it assesses whether a protein shows evidence of co-evolving with another, identifying the amino acids where this occurs. Finally, it investigates whether evidence of these relationships can inform structural predictions and the interpretation of such predictions. Not all steps in this framework have to be run, as results from each step can be enough to guide further experimentation, depending on the original inquiry.

We briefly revisit the test cases of the NAD and mTOR proteins to illustrate how the information flow feeds into different stages of this framework, and exemplify its intended use.

Paper 2 focused on PncA and Nampt in response to a broad inquiry into the level of evolution of the pathways involving these two enzymes, given that they act as precursor enzymes to two widely-recognised NAD salvage pathways. The mutually exclusive distribution pattern of PncA and Nampt across the tree of life could suggest an early divergence of these pathways, but this would need to be further verified in a genomic context. At this stage, evolutionary divergence cannot be concluded. This observation prompted a more focused investigation into potential additional characteristics matching the mutually exclusive pattern, exploring extremophilic or host-associated adaptations in bacteria and archaea. As PncA and Nampt are not suspected to interact, the analysis was confined to VisProPhyl, and was not carried further through the co-evolutionary pipeline or with AF2. Similarly, in the analysis of the taxonomic spread of the PRT family using VisProPhyl in Paper 1, while there were clear preferences for NAD enzymes within bacterial clades, indicating potential enzyme promiscuity, the emphasis was not on potential co-evolution of inter-enzymatic interactions when considering the broader context of the remaining PRTs. In these cases and also from two previous studies [Bockwoldt et al., 2019; Prentzell et al., 2021], VisProPhyl's phylogenetic overview provided enough information to direct the initial question towards more specific groups of organisms or selection criteria, or to motivate other experimental and computational work.

The mTOR proteins, on the other hand, provided an intriguing case to traverse the entire framework. Previous VisProPhyl analyses on mTORC1 proteins: G3BP1 and TSC2 [Prentzell et al., 2021], combined with repeat analyses with an updated database in a manuscript under preparation Heberle et al. from 2022 (data not included in thesis), identified G3BP1 and TSC2 as having a similar evolutionary timeframe. Experimental evidence had also made a case for the binding between the two proteins [Prentzell et al., 2021; Rehbein et al., 2021], although the exact site was not yet identified. These clues motivated a search for residue-level co-evolution between G3BP1 and TSC2, with the goal to narrow down the potential binding site regions to the domain or residue level, and the analysis moved onwards in the framework to co-evolutionary detection and structural prediction. One major question at this stage concerned how a binding site that co-evolves appears in terms of DCA scores arising from eukaryotic sequences, as DCA was developed for and benchmarked for prokaryotic sequences. Co-evolutionary analysis in our unpublished DCA work with Raptor and its substrates served to test for a co-evolutionary signature from regions known to bind, and highly suspected to co-evolve. No clear signature was identified through this test case, in terms of relative range of scores within which potential binding could be confidently detected, or a clustering pattern of co-evolutionary score ranges that could point out regions of interest. Co-evolutionary analysis of G3BP1:TSC2, the main targets, found a pattern between TSC2 C-ter regions, with two segments of G3BP1 (see 3.2, unpublished results), that are falling within the RRM and NTF2 domains. The need for structural comparison became apparent at this stage, and Paper 3 generated a series of AF2 and AF2M-predicted structures to compare. These structures brought forth the issue of the impact of disorder and conformational diversity on AF2M predictions. G3BP1 was predicted in the general vicinity of the TSC2 C-ter region, but not strongly interacting through either of its structured domains. Addition of binding partners confounded the predicted binding sites such that overall interpretability in line with previously calculated co-evolutionary scores was limited.

The application of this framework to the NAD and mTOR test cases in this thesis opened avenues to further refinements in the design and use of the individual components and the overall framework.

At the input stage, all three components are heavily MSA reliant, but vary in how stringently they allow for sequence inclusion, and therefore sample the available sequence space quite differently. VisProPhyl aims to sample as widely as possible the current scope of NCBI protein databases, and therefore its query step employs intentionally generous inclusion thresholds. Where it draws the line is strictly once cross-hits occur, or the same sequence hit being returned in separate queries of different proteins, which tends to only happen for proteins of high sequence similarity. The co-evolutionary workflow, on the other hand, requires a more stringent control on sequences included in the input MSA, to as closely meet the criteria of homologous, interacting sequences as possible. However, the organism space sampled cannot be too narrow, to allow for the occurrence of some level of sequence divergence. There is no consensus on the level of divergence necessary here, as this depends on the evolutionary trajectory of the target protein(s). Then, AF2 performs its own queries in several databases, creating multiple MSAs, but, similar to VisProPhyl, is aiming for inclusion and as much data as possible. Notably, AF2 can take an MSA as input, although this option was not used in the predictions of Paper 3. The information gained at each MSA construction step in this framework may be partly redundant, but there are portions that are unique to the query process or database that could provide feedback to other stages. A basic test of this would be to shuffle around the MSAs, such that, for example, VisProPhyl is run on the AF2 MSAs, co-evolutionary analysis is performed on the VisProPhyl MSA, and AF2 is run on a more limited MSA built in the co-evolutionary analysis step, for example. Then, comparison of the results would highlight, at least for the first two stages of the framework (AF2 results being a bit more opaque to input-based interpretation), what sequence overlap or sequence difference in the MSAs has highlighted in terms of protein distribution or co-evolutionary signal. In an ideal scenario, corrections could even be made at this stage, where certain organisms identified through the VisProPhyl query could be included in the co-evolutionary analysis MSA, for a more even sampling of the taxonomic tree, without getting too far from the clade containing the original target protein. Building in a module to collect, shuffle, and plug in the various MSAs within this framework would maximise the gain and leverage the utility of information from the various query strategies in this framework.

At the output stage, the framework would benefit from an additional software component responsible for the synthesis and visualisation of all three sets of results. Protein-level results (VisProPhyl) would be confined to one visualisation, and the residue-level results would be joined in another. On the front-end, this would be a user interface presenting side-by-side comparison of VisProPhyl tree and heatmap output with a second, joint window of AF2 predictions and co-evolutionary signal, perhaps in a parallel plot but with the option to highlight segments on the structure and compare predicted residue proximity with co-evolutionary signal. A joint visualisation of input MSA taxonomic characteristics would also be available from this component, complete with a clustering step to determine the effective number of sequences (see Fig. 11 for a potential representation of MSA visualisation). On the back-end, much of the code necessary for this component has been written, but individually for analysing results on separate occasions. The remaining work in this direction would involve constructing a workflow to manage file format compatibility and connect visualisation tools to output files. Essentially, the skeleton is there, and would be the ideal way to wrap up this framework, to push it one step further in accessibility, ease of use, and relevance to the target groups.

4.2 Navigating opportunities and challenges in analysing eukaryotic protein sequence data

With the exception of parts of the VisProPhyl analyses (in Paper 2 and a portion of Paper 1), the majority of this thesis focused on analysis of eukaryotic protein sequence data, and therefore

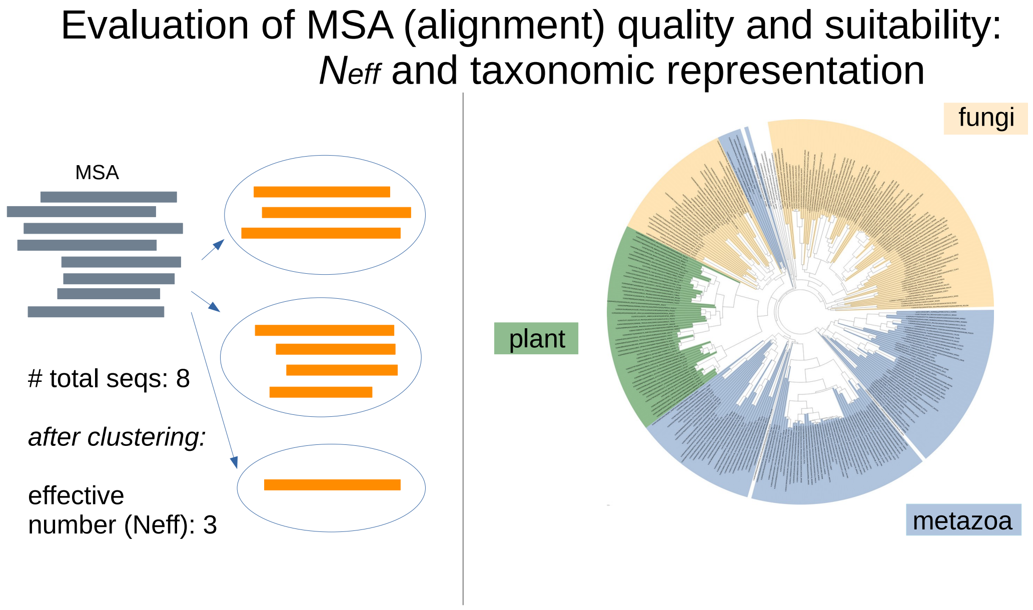
Evaluation of MSA (alignment) quality and suitability:
Neff and taxonomic representation
MSA
# total seqs: 8
after clustering:
effective number (Neff): 3
fungi
plant
metazoa

Figure 11: **MSA visualisation and evaluation component**: Proposed addition to the dual-axis framework, created over each input MSA. MSA sequences are clustered based on a consistent sequence similarity threshold, to allow for comparison of MSA depth. Sequences are visualised in a circular tree, and highlighted based on taxonomic group of interest.

repeatedly touched upon the unique opportunities and significant challenges that are resulting from the complexities of the eukaryotic genomic system.

At the most fundamental level, from the standpoint of sequence availability, this thesis work was possible due to a substantial increase of eukaryotic sequences in recent years, largely due to advancements in sequencing technologies, bioinformatics tools, and concurrent efforts to catalog proteomes of eukaryotic organisms. The current counts of eukaryotic sequences in UniProt, as of the March 2024 release on the EXPASY server [Gasteiger et al., 2003], sets the proportion to around 40%, compared to around 30% from a decade ago [Apweiler et al., 2004]. This puts the total eukaryotic sequence count at around 230,000. Looking at UniProt statistics as a reflection of overall eukaryotic sequence availability is akin to viewing the tip of the iceberg, in that growth in this gold standard, manually annotated set of protein sequences necessarily comes from an even more substantial growth in other databases that feed into it. For the methodology of this thesis framework to be effective, particularly for the training of the neural networks in AF2 structural prediction, it is essential to have a large number of sequences that encompass not only the major groups of eukaryotic organisms but also their lesser known relatives. One major opportunity of utilising eukaryotic protein sequence data, that this thesis revealed, is that the current sequence availability is of a workable scale for conducting sequence-heavy analyses like deep learning or co-evolutionary methods, and has a coverage that allows for evolutionary pattern extraction of tools like VisProPhyl to work effectively. Therefore with each year, a rerun of the framework on the same target proteins will very likely increase the accuracy of the results, assuming consistent growth in sampling.

Another advantage of analysing eukaryotic protein sequence data, highlighted by this thesis work, is a result of sequence availability, but from a functional annotation and comparative genomics motive. Focusing on eukaryotic protein sequences enables the identification and annotation of functional domains, motifs, and active sites that are more comparable for human biological and medical purposes, which is arguably the ultimate goal of much of the global volume of bioinformatic work. This is not at all the only application of the framework developed in this thesis, but definitely the inspiration for its development, and its intended use. The ability to perform deep, comparative analysis of protein sequences across different eukaryotic species provides insights into evolutionary relationships and the conservation of functional elements in relation to humans that before was not yet possible.

On the other hand, a recurrent challenge for this thesis work, arising from analysing eukaryotic protein data, concerns the complexity of eukaryotic genomes and the existence of paralogs. Eukaryotes have genomes which tend to be large, containing many repetitive sequences and introns. Eukaryotic genes often undergo duplication events, leading to multiple paralogs, or proteins resulting from duplicated genes that therefore share a common genetic origin. This complicates the accurate annotation and alignment of the resultant protein sequences, and adds a layer of complexity as each paralog has distinct functions and regulatory mechanisms associated with it. Section 3.2 makes this most obvious, as the co-evolutionary results from 4EBP1:Raptor and S6K:Raptor were expected to both pinpoint the TOS motif in the Raptor substrate proteins, but only did so for S6K:Raptor. The MSA depth of both analyses were similar, as was the taxonomic coverage, meaning that in terms of the general specifications of the sequence input set, 4EBP1:Raptor and S6K:Raptor were comparable. However, further clustering and background research into the TOR signaling role of the 4EBP paralogs vs. S6K paralogs, brought forth the theory that functionally, S6K paralogs were more homogeneous [Schalm and Blenis, 2002], whereas it was uncertain whether one of the 4EBPs, 4EBP3 still retained the same Raptor-binding function as 4EBP1, given that it lacks a N-ter RAIP motif that is necessary for the interaction [Tsukumo et al., 2016; Lee et al., 2008]. Inclusion of the paralogous sequence data in co-evolutionary analysis of S6K:Raptor then helped the analysis in that functional information was actually included. It is likely that inclusion of 4EBP paralog data confounded the co-evolutionary signal and hampered identification of the TOS binding region. It could also be that, since S6K has more paralogs, inclusion of their sequences presented a stronger co-evolutionary signal. Or, simply that the 4EBPs co-evolved less with Raptor. At this stage, one cannot know for sure, as this analysis would need to be scaled up to include more proteins for which enough paralog information exists. Less obvious, but still just as prominent of an issue, is that VisProPhyl presents an overview of a protein distribution that most likely also includes multiple paralogs, due to their sequence similarity to the target protein. One could attempt to set stricter inclusion/exclusion parameters at the query step, but this comes at the cost of meaningfully sampling a wider taxonomic group. In this way, a presence-absence pattern for a certain clade is normally not completely focused on the target protein, so the nuance of loss and gain of function across species for the group of target protein and its paralogs is not detectable. AF2 and AF2M are affected by the same need to construct a homologous (interacting) MSA, therefore the impact of paralogs is in effect for these predictors as well, although it is unclear exactly how. Take the case of 4EBP1 and 4EBP2, presented in Paper 3, for example. Perhaps if there is a bias in the PDB for a certain paralog (many studies focus on 4EBP1), which often occurs due simply to collective scientific bias, then AF2 is trained to represent this bias in its predictions, and no amount of padding the MSA with the actual target paralog will overcome this. Exclusion of templates in the model-building process of AF2 and AF2M is the best one can do to achieve a blank slate, but that does not override the training the algorithms were built on. 4EBP2 was experimentally shown to adopt various conformations, but AF2 and AF2M predictions seemed to cover this pattern. What this means from a usage perspective of the thesis framework is that unless paralog knowledge exists, the usability of the results is unclear. It is then necessary to supplement the results with an analysis on paralogs to the target protein, particularly to search for presence-absence at the functional domain level, which can involve returning to the input MSA and filtering through to discard unwanted paralog data.

Another challenging characteristic of protein sequences, though not unique to eukaryotes but significant due to its ubiquity when compared to prokaryotes [Macek et al., 2019], is the presence of post-translational modifications (PTMs) and its implications for analyses in this framework. Eukaryotic proteins frequently undergo PTMs, with phosphorylation being particularly relevant to the mTOR-associated proteins examined in this thesis. These modifications can significantly alter protein function and interactions, and a protein can contain multiple PTMs, many of which are dynamic and reversible. PTM data, however, is not yet integrated into the standard representation of a protein sequence, and also not easily incorporated into any of the methodologies presented in this thesis. If the previously mentioned paralog information were to be made more explicit and tunable within the framework, by the same logic, so should PTM information. Particular focus would be placed on evolution or co-evolution between modification sites and their flanking regions, for example, both of which have been shown to be under selective pressures [Bradley, 2022]. In this thesis, Paper 3 presented a phosphomimetic scheme to represent phosphorylated 4EBP, to test for change in structural prediction, but resulted in inconsistent results. This was most likely because the mimetic version of the sequence was similar enough to the original such that query results did not sample a set of sequences that could result in a structurally similar set to the phosphorylated state. As PTM databases continue to grow [Hornbeck et al., 2015], strengthening the framework to better integrate PTM information would be beneficial.

One final, practical consideration is that the size of eukaryotic protein complexes poses significant computational challenges, not just for this thesis but for any future work that rides the growing wave of eukaryotic sequence data availability. Although this is by far not a fixed pattern, the average size of protein complexes tends to be larger in eukaryotes compared to prokaryotes. This difference arises due to the greater complexity of eukaryotic cells, which often require more intricate regulatory mechanisms and exhibit greater functional diversity. Eukaryotic protein complexes often consist of numerous subunits; a classic example of this being the eukaryotic ribosome, with 80 proteins and 4 rRNA molecules (compared to the prokaryotic 55 proteins and 3 rRNA molecules), which reflects a more complex translational procedure. The mTORC1 complex is a 5-protein complex of 289KDa [Saxton and Sabatini, 2017], at a count of around 5000aa, which does not include all the regulatory partners. This is a scale that is problematic for sequence-based analyses that require querying, alignment, and deep learning in one overall computation - the resources simply are not available (at least to the Norwegian scientific community) at this stage. To add to this, bioinformatic sequence-based tools that were developed with the the smaller prokaryotic system in mind, such as the co-evolutionary methodology, placed less focus on optimisation and parallelisation. In this thesis, as long as the total residue count stayed under a comfortable 2000aa, computation was not an issue. The TSC2 and G3BP1 dimer surpassed this. AF2M structural modeling of the dimer and overall complex (see Paper 3), reached the time and memory limit, and actually for AF2M never fully completed through all the recycling rounds and AMBER relaxation. For structural modeling, adopting a more piecewise approach, either truncating the complex to regions of interest (what was done in Paper 3 with the TSC complex), or modeling full proteins but only portions of the complex, and docking the predictions, could be viable workarounds. For DCA, pre-processing of sequences to reduce the overall MSA size (e.g. clustering), or experimenting with more efficient numerical solvers of the inference step, or rewriting of the code for GPUs could boost efficiency. It would be worth embedding such optimisation tactics into the framework proposed in this thesis, but ultimately, the push for optimisation has to come from the developers of the external software used in the framework.

4.3 Understanding the interpretability of results in a broader context

Issues of interpretability frequently arose when assessing the various results of this framework, particularly when taking into consideration the incomplete context in which the results are derived.

The protein sequence space available to bioinformatics work is, however large, still incomplete. Biases exist in the databases due to general scientific focus, funding, and inherent biological factors. A significant proportion of sequence data comes from a few well-studied model organisms. This results in an overrepresentation of these organisms and underrepresentation of many other eukaryotic species, especially those from understudied habitats or with no direct economic or medical importance, creating gaps in the database. There is a temporal bias as well, as recently sequenced organisms are more likely to benefit from advanced sequencing technologies, resulting in better quality data compared to older sequences obtained using outdated methods. This means that a presence-absence distribution of proteins in a taxonomic clade, as extracted by

VisProPhyl, can be a reflection of the true distribution pattern, but is seen through a lens of all the biases layered on top. In reconstructing the evolutionary trajectory of an enzyme, we cannot know for sure that an organism has lost it, but it becomes more likely if relatives all do not have it. Even the assignment of presence of an enzyme to an organism depends on the significance level of the query parameters, as it is not practically feasible to go through every organism and check for annotation correctness. The takeaway here is that when concluding co-evolutionary or evolutionary results from VisProPhyl's overview, one should be cautious and at the very least aware of the general extent to which target proteins have been studied, if there are many verified sequences, what is known about paralogs, and how this may impact the evaluation of the output.

Then, regarding co-evolutionary results, a frequent roadblock to interpretability arose from common practices surrounding the co-evolutionary coupling scores. As mentioned before, scores are not inter-comparable across co-evolutionary analyses of different proteins or protein complexes. One study identified a score threshold of 0.6 as a meaningful bar for bacterial ribosomal proteins above which co-variation and potential co-evolution reliably indicated residue proximity Ovchinnikov et al. [2014]. This does not necessarily apply across different bacterial complexes Si et al. [2022], and currently, no such level has been established for the eukaryotic case. Then, typically, a subset of top-scoring residue pairs is selected, and their positions are analyzed to infer structural contacts. This approach has helped in predicting the tertiary structure of proteins, particularly for proteins with unknown structures or those that are difficult to crystallize. This information is then used to feed into methods like molecular dynamics simulations or machine learning models. However, there is in fact a whole spectrum of scores generated through these methods, much of which is typically ignored. Are low-scoring residue pairs necessarily highly conserved? What about mid-range scores of residue pairs? Our attempts to map conservation to co-evolutionary scores in the parallel plots of Section 3.2 show that it is not always the case that highly variable regions are co-evolving, which makes sense from an independent evolutionary standpoint, but that sometimes, highly co-evolving residue pairs are not highly variable. Using co-evolutionary methods to accurately identify non-co-evolving regions would be an excellent use of this method, but a lot more work is needed in this direction.

Perhaps the biggest question of interpretability was raised by the results of Paper 3. The interpretability of AF2 and AF2M predictions varies across different regions of the protein structure, fluctuating between high-confidence and low-confidence regions. Regions with high pLDDT scores (typically above 70-80) are considered high-confidence regions, often aligning closely with experimentally determined structures that are stable and structured. Regions with low pLDDT scores (below 50-60), are considered low-confidence regions, corresponding to disordered or flexible parts of the protein, such as loops, termini, or intrinsically disordered regions (IDRs), and in our case represented a significant proportion of the known functional sites of each protein and protein complex we modeled. Low confidence regions indeed often aligned with known IDR regions, but experimental evidence pointed towards, at least for the 4EBP:EIF4E case, transient adoption of structure upon binding. Attempts to tease out a structural prediction in these regions, to get any idea of how a structure may be, were largely unsuccessful. Additionally, structural variation was observed in the various models produced by AF2 and AF2M, but the documentation of AF2 and AF2M clearly warns against interpreting low confidence regions, and also against drawing any sort of structural consensus from the model variation Jumper et al. [2021]. This calls into question the usability of AF2 for these cases, but also whether its training dataset, the PDB itself, is an comprehensive enough reference to train a robust predictor of protein and protein complex structure. The PDB's static picture of protein structure influences the tendencies of AF2, therefore making it unsuitable for modeling much of the dynamics that are necessary for PPI. AF2 becomes, in essence, a predictor of whether a protein can crystallise, as the majority of structures on the PDB (around 74.5%, see `https://www.rcsb.org/stats/summary`) are resolved with X-ray crystallography. Tendencies for AF2M to predict symmetry Zhu et al. [2023], and its limitations to fewer chains [Bryant et al., 2022] also constrict predictive possibility further. Increases in the number of nuclear magnetic resonance (NMR) spectroscopy and cryo-electron microscopy (cryo-EM) structures could present a more variable set for re-training of the neural networks modules in AF2. It could be that AF2 can capture some of the dynamics, and stores this in the variation of its output, but without a proper reference, one cannot investigate this further. As this thesis was being written, however, AlphaFold3 [Abramson et al., 2024] was released, promising to model biomolecular interactions, so perhaps the future is here, and interpretability will get a much-needed boost in the dynamics direction.

5 Conclusion and outlook

This thesis developed a comprehensive framework for analysing protein evolutionary and co-evolutionary patterns. This computational framework was built over three studies and additional work that addressed two major goals: inferring pathway diversity and organism-specific preferences through enzyme presence-absence patterns, and predicting PPIs and protein structures across proteins of varying disorder, size, and complexity. The framework allows for explicit or implicit calculations of evolution or co-evolution, at both residue and whole-protein levels, to leverage the information gained from their combination. VisProPhyl provided explicit, protein-level analyses, while the co-evolutionary pipeline (DCA) and AlphaFold2 (AF2) offered residue-level analyses. Applying this framework to NAD and mTOR proteins demonstrated its utility and highlighted challenges related to protein disorder and conformational diversity, and, from a more general perspective, demonstrated that while eukaryotic sequence data offers significant research opportunities, challenges remain in handling genomic complexity and improving result interpretability. Future improvements will enhance the framework by refining input and output stages and incorporating state-of-the-art structural predictors like AlphaFold3. The joint reliance on MSA input underscored the importance of sequence selection criteria, and allowing for the shuffling of MSAs across the different analyses methods can maximise the utility of information from different query strategies. Integrating post-query information on existence of paralogs and PTM sites will allow the user to have a better overview of the input to the framework, and tune the input according to the desired MSA representation of the targets. Enhancements in visualization and result integration will make these methodologies more accessible, contributing to a deeper understanding of protein interactions and functions. Optimising the framework by reorganizing it into querying, processing, and analysis modules will streamline the computational workflow, making the analysis more efficient.

Bibliography

Database resources of the National Center for Biotechnology Information. *Nucleic acids research*, 44(D1):D7–19, Jan. 2016. ISSN 1362-4962 0305-1048. doi: 10.1093/nar/gkv1290.

J. Abramson, J. Adler, J. Dunger, R. Evans, T. Green, A. Pritzel, O. Ronneberger, L. Willmore, A. J. Ballard, J. Bambrick, S. W. Bodenstein, D. A. Evans, C.-C. Hung, M. O'Neill, D. Reiman, K. Tunyasuvunakool, Z. Wu, A. Žemgulytė, E. Arvaniti, C. Beattie, O. Bertolli, A. Bridgland, A. Cherepanov, M. Congreve, A. I. Cowen-Rivers, A. Cowie, M. Figurnov, F. B. Fuchs, H. Gladman, R. Jain, Y. A. Khan, C. M. R. Low, K. Perlin, A. Potapenko, P. Savy, S. Singh, A. Stecula, A. Thillaisundaram, C. Tong, S. Yakneen, E. D. Zhong, M. Zielinski, A. Žídek, V. Bapst, P. Kohli, M. Jaderberg, D. Hassabis, and J. M. Jumper. Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature*, May 2024. ISSN 1476-4687. doi: 10.1038/s41586-024-07487-w.

G. E. Allen. Mendel and modern genetics: The legacy for today. *Endeavour*, 27(2):63–68, June 2003. ISSN 0160-9327. doi: 10.1016/S0160-9327(03)00065-6.

S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. Basic local alignment search tool. *Journal of Molecular Biology*, 215(3):403–410, Oct. 1990. ISSN 0022-2836. doi: 10.1016/S0022-2836(05)80360-2.

S. Amjad, S. Nisar, A. A. Bhat, A. R. Shah, M. P. Frenneaux, K. Fakhro, M. Haris, R. Reddy, Z. Patay, J. Baur, and P. Bagga. Role of NAD(+) in regulating cellular and metabolic signaling pathways. *Molecular metabolism*, 49:101195, July 2021. ISSN 2212-8778. doi: 10.1016/j.molmet.2021.101195.

R. Apweiler, A. Bairoch, C. H. Wu, W. C. Barker, B. Boeckmann, S. Ferro, E. Gasteiger, H. Huang, R. Lopez, M. Magrane, M. J. Martin, D. A. Natale, C. O'Donovan, N. Redaschi, and L.-S. L. Yeh. UniProt: The Universal Protein knowledgebase. *Nucleic acids research*, 32 (Database issue):D115–119, Jan. 2004. ISSN 1362-4962 0305-1048. doi: 10.1093/nar/gkh131.

H. Ashkenazy, S. Abadi, E. Martz, O. Chay, I. Mayrose, T. Pupko, and N. Ben-Tal. ConSurf 2016: An improved methodology to estimate and visualize evolutionary conservation in macromolecules. *Nucleic Acids Research*, 44(W1):W344–W350, July 2016. ISSN 0305-1048. doi: 10.1093/nar/gkw408.

C. Baldassi, M. Zamparo, C. Feinauer, A. Procaccini, R. Zecchina, M. Weigt, and A. Pagnani. Fast and accurate multivariate Gaussian modeling of protein families: Predicting residue contacts and protein-interaction partners. *PloS one*, 9(3):e92721, 2014. ISSN 1932-6203. doi: 10.1371/journal.pone.0092721.

K. R. Benson. T. H. Morgan's resistance to the chromosome theory. *Nature Reviews Genetics*, 2(6):469–474, June 2001. ISSN 1471-0064. doi: 10.1038/35076532.

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 Research*, 28(1):235–242, Jan. 2000. ISSN 0305-1048. doi: 10.1093/nar/28.1.235.

M. Bockwoldt, D. Houry, M. Niere, T. I. Gossmann, I. Reinartz, A. Schug, M. Ziegler, and I. Heiland. Identification of evolutionary and kinetic drivers of NAD-dependent signaling. *Proceedings of the National Academy of Sciences*, 116(32):15957–15966, Aug. 2019. doi: 10.1073/pnas.1902346116.

R. Böhm, S. Imseng, R. P. Jakob, M. N. Hall, T. Maier, and S. Hiller. The dynamic mechanism of 4E-BP1 recognition and phosphorylation by mTORC1. *Molecular Cell*, 81(11):2403–2416.e5, June 2021. ISSN 1097-2765. doi: 10.1016/j.molcel.2021.03.031.

D. Bradley. The evolution of post-translational modifications. *Current Opinion in Genetics & Development*, 76:101956, Oct. 2022. ISSN 0959-437X. doi: 10.1016/j.gde.2022.101956.

P. Bryant, G. Pozzati, W. Zhu, A. Shenoy, P. Kundrotas, and A. Elofsson. Predicting the structure of large protein complexes using AlphaFold and Monte Carlo tree search. *Nature Communications*, 13(1):6028, Oct. 2022. ISSN 2041-1723. doi: 10.1038/s41467-022-33729-4.

F. Buhr, S. Jha, M. Thommen, J. Mittelstaet, F. Kutz, H. Schwalbe, M. V. Rodnina, and A. A. Komar. Synonymous Codons Direct Cotranslational Folding toward Different Protein Conformations. *Molecular cell*, 61(3):341–351, Feb. 2016. ISSN 1097-4164 1097-2765. doi: 10.1016/j.molcel.2016.01.008.

C. Camacho, G. Coulouris, V. Avagyan, N. Ma, J. Papadopoulos, K. Bealer, and T. L. Madden. BLAST+: Architecture and applications. *BMC Bioinformatics*, 10(1):421, Dec. 2009. ISSN 1471-2105. doi: 10.1186/1471-2105-10-421.

D. Carmona, C. R. Fitzpatrick, and M. T. J. Johnson. Fifty years of co-evolution and beyond: Integrating co-evolution from molecules to species. *Molecular Ecology*, 24(21):5315–5329, Nov. 2015. ISSN 0962-1083. doi: 10.1111/mec.13389.

E. Caron, S. Ghosh, Y. Matsuoka, D. Ashton-Beaucage, M. Therrien, S. Lemieux, C. Perreault, P. P. Roux, and H. Kitano. A comprehensive map of the mTOR signaling network. *Molecular systems biology*, 6:453, Dec. 2010. ISSN 1744-4292. doi: 10.1038/msb.2010.108.

A. Chiarugi, C. Dölle, R. Felici, and M. Ziegler. The NAD metabolome — a key determinant of cancer cell biology. *Nature Reviews Cancer*, 12(11):741–752, Nov. 2012. ISSN 1474-1768. doi: 10.1038/nrc3340.

F. Crick. The origin of the genetic code. *Journal of Molecular Biology*, 38(3):367–379, Dec. 1968. ISSN 0022-2836. doi: 10.1016/0022-2836(68)90392-6.

G. Csárdi and T. Nepusz. The igraph software package for complex network research. 2006.

C. Darwin. On the various contrivances by which British and foreign orchids are fertilized. *Murray, London*, 365, 1862.

C. Darwin, 1809-1882. *On the Origin of Species by Means of Natural Selection, or Preservation of Favoured Races in the Struggle for Life*. London : John Murray, 1859, 1859.

D. de Juan, F. Pazos, and A. Valencia. Emerging methods in protein co-evolution. *Nature Reviews Genetics*, 14(4):249–261, Apr. 2013. ISSN 1471-0064. doi: 10.1038/nrg3414.

Y. Ding, X. Li, G. P. Horsman, P. Li, M. Wang, J. Li, Z. Zhang, W. Liu, B. Wu, Y. Tao, and Y. Chen. Construction of an Alternative NAD(+) De Novo Biosynthesis Pathway. *Advanced science (Weinheim, Baden-Wurttemberg, Germany)*, 8(9):2004632, May 2021. ISSN 2198-3844. doi: 10.1002/advs.202004632.

T. Dou, C. Ji, S. Gu, J. Xu, J. Xu, K. Ying, Y. Xie, and Y. Mao. Co-evolutionary analysis of insulin/insulin like growth factor 1 signal pathway in vertebrate species. *Frontiers in bioscience : a journal and virtual library*, 11:380–388, Jan. 2006. ISSN 1093-9946 1093-4715. doi: 10.2741/1805.

S. Dunn, L. Wahl, and G. Gloor. Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction. *Bioinformatics*, 24(3):333–340, Feb. 2008. ISSN 1367-4803. doi: 10.1093/bioinformatics/btm604.

R. C. Edgar. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic acids research*, 32(5):1792–1797, 2004. ISSN 1362-4962 0305-1048. doi: 10.1093/nar/gkh340.

PR. Ehrlich and PH. Raven. Butterflies and plants - a study in coevolutoin. *EVOLUTION*, 18 (4):586–608, 1964. ISSN 0014-3820. doi: 10.2307/2406212.

M. Ekeberg, C. Lövkvist, Y. Lan, M. Weigt, and E. Aurell. Improved contact prediction in proteins: Using pseudolikelihoods to infer Potts models. *Physical Review E*, 87(1):012707, Jan. 2013. doi: 10.1103/PhysRevE.87.012707.

R. Evans, M. O'Neill, A. Pritzel, N. Antropova, A. Senior, T. Green, A. Žídek, R. Bates, S. Blackwell, J. Yim, O. Ronneberger, S. Bodenstein, M. Zielinski, A. Bridgland, A. Potapenko, A. Cowie, K. Tunyasuvunakool, R. Jain, E. Clancy, P. Kohli, J. Jumper, and D. Hassabis. Protein complex prediction with AlphaFold-Multimer. *bioRxiv : the preprint server for biology*, 2022. doi: 10.1101/2021.10.04.463034.

R. D. Finn, J. Clements, and S. R. Eddy. HMMER web server: Interactive sequence similarity searching. *Nucleic Acids Research*, 39(suppl_2):W29–W37, July 2011. ISSN 0305-1048. doi: 10.1093/nar/gkr367.

A. A. Fodor and R. W. Aldrich. Influence of conservation on calculations of amino acid covariance in multiple sequence alignments. *Proteins: Structure, Function, and Bioinformatics*, 56 (2):211–221, Aug. 2004. ISSN 0887-3585. doi: 10.1002/prot.20098.

R. E. FRANKLIN and R. G. GOSLING. Molecular Configuration in Sodium Thymonucleate. *Nature*, 171(4356):740–741, Apr. 1953. ISSN 1476-4687. doi: 10.1038/171740a0.

E. Gasteiger, A. Gattiker, C. Hoogland, I. Ivanyi, R. D. Appel, and A. Bairoch. ExPASy: The proteomics server for in-depth protein knowledge and analysis. *Nucleic acids research*, 31 (13):3784–3788, July 2003. ISSN 1362-4962 0305-1048. doi: 10.1093/nar/gkg563.

F. Gazzaniga, R. Stebbins, S. Z. Chang, M. A. McPeek, and C. Brenner. Microbial NAD metabolism: Lessons from comparative genomics. *Microbiology and molecular biology reviews : MMBR*, 73(3):529–541, Table of Contents, Sept. 2009. ISSN 1098-5557 1092-2172. doi: 10.1128/MMBR.00042-08.

T. I. Gossmann, M. Ziegler, P. Puntervoll, L. F. de Figueiredo, S. Schuster, and I. Heiland. NAD(+) biosynthesis and salvage–a phylogenetic perspective. *The FEBS journal*, 279(18): 3355–3363, Sept. 2012. ISSN 1742-4658 1742-464X. doi: 10.1111/j.1742-4658.2012.08559.x.

L. Hakes, S. C. Lovell, S. G. Oliver, and D. L. Robertson. Specificity in protein interactions and its relationship with sequence diversity and coevolution. *Proceedings of the National Academy of Sciences*, 104(19):7999–8004, May 2007. doi: 10.1073/pnas.0609962104.

G. Hamoir. The discovery of meiosis by E. Van Beneden, a breakthrough in the morphological phase of heredity. *The International journal of developmental biology*, 36(1):9–15, Mar. 1992. ISSN 0214-6282.

P. V. Hornbeck, B. Zhang, B. Murray, J. M. Kornhauser, V. Latham, and E. Skrzypek. PhosphoSitePlus, 2014: Mutations, PTMs and recalibrations. *Nucleic acids research*, 43(Database issue):D512–520, Jan. 2015. ISSN 1362-4962 0305-1048. doi: 10.1093/nar/gku1267.

J. Huang and B. D. Manning. The TSC1-TSC2 complex: A molecular switchboard controlling cell growth. *The Biochemical journal*, 412(2):179–190, June 2008. ISSN 1470-8728 0264-6021. doi: 10.1042/BJ20080281.

J. Huerta-Cepas, F. Serra, and P. Bork. ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data. *Molecular Biology and Evolution*, 33(6):1635–1638, June 2016. ISSN 0737-4038. doi: 10.1093/molbev/msw046.

J. Hunter. Matplotlib: A 2D Graphics Environment. *Computing in Science & Engineering*, 9: 90–95, June 2007. doi: 10.1109/MCSE.2007.55.

P. T. Inc. Collaborative data science. https://plot.ly, 2015.

S. D. Johnson and B. Anderson. Coevolution Between Food-Rewarding Flowers and Their Pollinators. *Evolution: Education and Outreach*, 3(1):32–39, Mar. 2010. ISSN 1936-6434. doi: 10.1007/s12052-009-0192-6.

D. Juan, F. Pazos, and A. Valencia. High-confidence prediction of global interactomes based on genome-wide coevolutionary networks. *Proceedings of the National Academy of Sciences*, 105(3):934–939, Jan. 2008. doi: 10.1073/pnas.0709671105.

J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, A. Bridgland, C. Meyer, S. A. A. Kohl, A. J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A. W. Senior, K. Kavukcuoglu, P. Kohli, and D. Hassabis. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589, Aug. 2021. ISSN 1476-4687. doi: 10.1038/s41586-021-03819-2.

K. Katoh, K. Misawa, K.-i. Kuma, and T. Miyata. MAFFT: A novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research*, 30(14):3059–3066, July 2002. ISSN 0305-1048. doi: 10.1093/nar/gkf436.

D.-H. Kim, D. Sarbassov, S. Ali, J. King, R. Latek, H. Erdjument-Bromage, P. Tempst, and D. Sabatini. mTOR Interacts with Raptor to Form a Nutrient-Sensitive Complex that Signals to the Cell Growth Machinery. *Cell*, 110:163–75, Aug. 2002. doi: 10.1016/S0092-8674(02)00808-5.

M. Kimura. Evolutionary Rate at the Molecular Level. *Nature*, 217(5129):624–626, Feb. 1968. ISSN 1476-4687. doi: 10.1038/217624a0.

J. L. King and T. H. Jukes. Non-Darwinian Evolution. *Science*, 164(3881):788–798, May 1969. doi: 10.1126/science.164.3881.788.

M. Laplante and D. M. Sabatini. mTOR signaling at a glance. *Journal of cell science*, 122(Pt 20):3589–3594, Oct. 2009. ISSN 1477-9137 0021-9533. doi: 10.1242/jcs.051011.

V. H. Y. Lee, T. Healy, B. D. Fonseca, A. Hayashi, and C. G. Proud. Analysis of the regulatory motifs in eukaryotic initiation factor 4E-binding protein 1. *The FEBS Journal*, 275(9):2185–2199, May 2008. ISSN 1742-464X. doi: 10.1111/j.1742-4658.2008.06372.x.

S. C. Lovell and D. L. Robertson. An Integrated View of Molecular Coevolution in Protein–Protein Interactions. *Molecular Biology and Evolution*, 27(11):2567–2575, Nov. 2010. ISSN 0737-4038. doi: 10.1093/molbev/msq144.

M. Lynch. Mutation pressure, drift, and the pace of molecular coevolution. *Proceedings of the National Academy of Sciences*, 120(27):e2306741120, July 2023. doi: 10.1073/pnas.2306741120.

B. Macek, K. Forchhammer, J. Hardouin, E. Weber-Ban, C. Grangeasse, and I. Mijakovic. Protein post-translational modifications in bacteria. *Nature Reviews Microbiology*, 17(11):651–664, Nov. 2019. ISSN 1740-1534. doi: 10.1038/s41579-019-0243-0.

L. C. Martin, G. B. Gloor, S. D. Dunn, and L. M. Wahl. Using information theory to search for co-evolving residues in proteins. *Bioinformatics (Oxford, England)*, 21(22):4116–4124, Nov. 2005. ISSN 1367-4803. doi: 10.1093/bioinformatics/bti671.

I. Mayrose, D. Graur, N. Ben-Tal, and T. Pupko. Comparison of Site-Specific Rate-Inference Methods for Protein Sequences: Empirical Bayesian Methods Are Superior. *Molecular Biology and Evolution*, 21(9):1781–1791, Sept. 2004. ISSN 0737-4038. doi: 10.1093/molbev/msh194.

F. Morcos, A. Pagnani, B. Lunt, A. Bertolino, D. S. Marks, C. Sander, R. Zecchina, J. N. Onuchic, T. Hwa, and M. Weigt. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proceedings of the National Academy of Sciences of the United States of America*, 108(49):E1293–1301, Dec. 2011. ISSN 1091-6490 0027-8424. doi: 10.1073/pnas.1111471108.

S. Ovchinnikov, H. Kamisetty, and D. Baker. Robust and accurate prediction of residue–residue interactions across protein interfaces using evolutionary information. *eLife*, 3:e02030, May 2014. ISSN 2050-084X. doi: 10.7554/eLife.02030.

A. Pauw, J. Stofberg, and R. J. Waterman. Flies and flowers in Darwin's race. *Evolution*, 63 (1):268–279, Jan. 2009. ISSN 0014-3820. doi: 10.1111/j.1558-5646.2008.00547.x.

N. Paweletz. Walther Flemming: Pioneer of mitosis research. *Nature reviews. Molecular cell biology*, 2(1):72–75, Jan. 2001. ISSN 1471-0072. doi: 10.1038/35048077.

F. Pazos and A. Valencia. Similarity of phylogenetic trees as indicator of protein-protein interaction. *Protein engineering*, 14(9):609–614, Sept. 2001. ISSN 0269-2139. doi: 10.1093/protein/14.9.609.

F. Pazos and A. Valencia. Protein co-evolution, co-adaptation and interactions. *The EMBO Journal*, 27(20):2648–2655, Oct. 2008. ISSN 0261-4189. doi: 10.1038/emboj.2008.189.

M. Pellegrini, E. M. Marcotte, M. J. Thompson, D. Eisenberg, and T. O. Yeates. Assigning protein functions by comparative genome analysis: Protein phylogenetic profiles. *Proceedings of the National Academy of Sciences of the United States of America*, 96(8):4285–4288, Apr. 1999. ISSN 0027-8424 1091-6490. doi: 10.1073/pnas.96.8.4285.

J. Preiss and P. Handler. Biosynthesis of diphosphopyridine nucleotide. I. Identification of intermediates. *The Journal of biological chemistry*, 233(2):488–492, Aug. 1958. ISSN 0021-9258.

M. T. Prentzell, U. Rehbein, M. Cadena Sandoval, A.-S. De Meulemeester, R. Baumeister, L. Brohée, B. Berdel, M. Bockwoldt, B. Carroll, S. R. Chowdhury, A. von Deimling, C. Demetriades, G. Figlia, M. E. G. de Araujo, A. M. Heberle, I. Heiland, B. Holzwarth, L. A. Huber, J. Jaworski, M. Kedra, K. Kern, A. Kopach, V. I. Korolchuk, I. van 't Land-Kuper, M. Macias, M. Nellist, W. Palm, S. Pusch, J. M. Ramos Pittol, M. Reil, A. Reintjes, F. Reuter,

J. R. Sampson, C. Scheldeman, A. Siekierska, E. Stefan, A. A. Teleman, L. E. Thomas, O. Torres-Quesada, S. Trump, H. D. West, P. de Witte, S. Woltering, T. E. Yordanov, J. Zmorzynska, C. A. Opitz, and K. Thedieck. G3BPs tether the TSC complex to lysosomes and suppress mTORC1 signaling. *Cell*, 184(3):655–674.e27, 2021. ISSN 0092-8674. doi: 10.1016/j.cell.2020.12.024.

A. K. Ramani and E. M. Marcotte. Exploiting the co-evolution of interacting proteins to discover interaction specificity. *Journal of molecular biology*, 327(1):273–284, Mar. 2003. ISSN 0022-2836. doi: 10.1016/s0022-2836(03)00114-1.

U. Rehbein, M. T. Prentzell, M. Cadena Sandoval, A. M. Heberle, E. P. Henske, C. A. Opitz, and K. Thedieck. The TSC complex-mTORC1 axis: From lysosomes to stress granules and back. *Frontiers in Cell and Developmental Biology*, 9, 2021. ISSN 2296-634X. doi: 10.3389/fcell.2021.751892.

J. P. Roney and S. Ovchinnikov. State-of-the-Art Estimation of Protein Model Accuracy Using AlphaFold. *Physical Review Letters*, 129(23):238101, Nov. 2022. doi: 10.1103/PhysRevLett.129.238101.

S. Ruggieri, G. Orsomando, L. Sorci, and N. Raffaelli. Regulation of NAD biosynthetic enzymes modulates NAD-sensing processes to shape mammalian cell physiology under varying biological cues. *Cofactor-dependent proteins: evolution, chemical diversity and bio-applications*, 1854(9):1138–1149, Sept. 2015. ISSN 1570-9639. doi: 10.1016/j.bbapap.2015.02.021.

R. A. Saxton and D. M. Sabatini. mTOR Signaling in Growth, Metabolism, and Disease. *Cell*, 168(6):960–976, Mar. 2017. ISSN 1097-4172 0092-8674. doi: 10.1016/j.cell.2017.02.004.

S. S. Schalm and J. Blenis. Identification of a Conserved Motif Required for mTOR Signaling. *Current Biology*, 12(8):632–639, Apr. 2002. ISSN 0960-9822. doi: 10.1016/S0960-9822(02)00762-5.

C. L. Schoch, S. Ciufo, M. Domrachev, C. L. Hotton, S. Kannan, R. Khovanskaya, D. Leipe, R. Mcveigh, K. O'Neill, B. Robbertse, S. Sharma, V. Soussov, J. P. Sullivan, L. Sun, S. Turner, and I. Karsch-Mizrachi. NCBI Taxonomy: A comprehensive update on curation, resources and tools. *Database : the journal of biological databases and curation*, 2020, Jan. 2020. ISSN 1758-0463. doi: 10.1093/database/baaa062.

Schrödinger, LLC. The PyMOL molecular graphics system, version 1.8. Nov. 2015.

A. Schug, M. Weigt, J. N. Onuchic, T. Hwa, and H. Szurmant. High-resolution protein complexes from integrating genomic information with molecular simulation. *Proceedings of the National Academy of Sciences*, 106(52):22124–22129, Dec. 2009. doi: 10.1073/pnas.0912100106.

X. Shen, S. Song, C. Li, and J. Zhang. Synonymous mutations in representative yeast genes are mostly strongly non-neutral. *Nature*, 606(7915):725–731, June 2022. ISSN 1476-4687 0028-0836. doi: 10.1038/s41586-022-04823-w.

Y. Si, Y. Zhang, and C. Yan. A reproducibility analysis-based statistical framework for residue–residue evolutionary coupling detection. *Briefings in Bioinformatics*, 23(2):bbab576, Mar. 2022. ISSN 1477-4054. doi: 10.1093/bib/bbab576.

G. M. Süel, S. W. Lockless, M. A. Wall, and R. Ranganathan. Evolutionarily conserved networks of residues mediate allosteric communication in proteins. *Nature Structural Biology*, 10(1): 59–69, Jan. 2003. ISSN 1545-9985. doi: 10.1038/nsb881.

P. H. Thrall and J. J. Burdon. Evolution of virulence in a plant host-pathogen metapopulation. *Science*, 299(5613):1735–7, Mar. 2003. ISSN 00368075.

Y. Tsukumo, T. Alain, B. D. Fonseca, R. Nadon, and N. Sonenberg. Translation control during prolonged mTORC1 inhibition mediated by 4E-BP3. *Nature Communications*, 7(1):11776, June 2016. ISSN 2041-1723. doi: 10.1038/ncomms11776.

G. Uguzzoni, S. John Lovis, F. Oteri, A. Schug, H. Szurmant, and M. Weigt. Large-scale identification of coevolution signals across homo-oligomeric protein interfaces by direct coupling analysis. *Proceedings of the National Academy of Sciences*, 114(13):E2662–E2671, Mar. 2017. doi: 10.1073/pnas.1615068114.

J. D. WATSON and F. H. C. CRICK. Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. *Nature*, 171(4356):737–738, Apr. 1953. ISSN 1476-4687. doi: 10.1038/171737a0.

M. Weigt, R. A. White, H. Szurmant, J. A. Hoch, and T. Hwa. Identification of direct residue contacts in protein–protein interaction by message passing. *Proceedings of the National Academy of Sciences*, 106(1):67–72, Jan. 2009. doi: 10.1073/pnas.0805923106.

H. Yang, X. Jiang, B. Li, H. J. Yang, M. Miller, A. Yang, A. Dhar, and N. P. Pavletich. Mechanisms of mTORC1 activation by RHEB and inhibition by PRAS40. *Nature*, 552(7685): 368–373, Dec. 2017. ISSN 1476-4687. doi: 10.1038/nature25023.

C.-H. Yeang and D. Haussler. Detecting Coevolution in and among Protein Domains. *PLOS Computational Biology*, 3(11):e211, Nov. 2007. doi: 10.1371/journal.pcbi.0030211.

M. B. Zerihun, F. Pucci, E. K. Peter, and A. Schug. Pydca v1.0: A comprehensive software for direct coupling analysis of RNA and protein sequences. *Bioinformatics*, 36(7):2264–2265, Apr. 2020. ISSN 1367-4803. doi: 10.1093/bioinformatics/btz892.

J. Zhang and J.-R. Yang. Determinants of the rate of protein sequence evolution. *Nature Reviews Genetics*, 16(7):409–420, July 2015. ISSN 1471-0064. doi: 10.1038/nrg3950.

H. Zhou and E. Jakobsson. Predicting Protein-Protein Interaction by the Mirrortree Method: Possibilities and Limitations. *PLOS ONE*, 8(12):e81100, Dec. 2013. doi: 10.1371/journal.pone.0081100.

Z. Zhou, Y. Dang, M. Zhou, L. Li, C.-H. Yu, J. Fu, S. Chen, and Y. Liu. Codon usage is an important determinant of gene expression levels largely through its effects on transcription. *Proceedings of the National Academy of Sciences of the United States of America*, 113(41): E6117–E6125, Oct. 2016. ISSN 1091-6490 0027-8424. doi: 10.1073/pnas.1606724113.

W. Zhu, A. Shenoy, P. Kundrotas, and A. Elofsson. Evaluation of AlphaFold-Multimer prediction on multi-chain protein complexes. *Bioinformatics*, 39(7):btad424, July 2023. ISSN 1367-4811. doi: 10.1093/bioinformatics/btad424.

M. Ziegler and M. Niere. NAD+ surfaces again. *The Biochemical journal*, 382(Pt 3):e5–6, Sept. 2004. ISSN 1470-8728 0264-6021. doi: 10.1042/BJ20041217.

Z. Zou, T. Tao, H. Li, and X. Zhu. mTOR signaling pathway and mTOR inhibitors in cancer: Progress and challenges. *Cell & Bioscience*, 10(1):31, Mar. 2020. ISSN 2045-3701. doi: 10.1186/s13578-020-00396-1.

6 Papers 1-3

metabolites

Article

Early Evolutionary Selection of NAD Biosynthesis Pathway in Bacteria

Suraj Sharma [1], Yin-Chen Hsieh [1], Jörn Dietze [1], Mathias Bockwoldt [2], Øyvind Strømland [3], Mathias Ziegler [3] and Ines Heiland [1,4,*]

[1] Department of Arctic and Marine Biology, Faculty of Biosciences, Fisheries and Economics, UiT The Arctic University of Norway, 9037 Tromsø, Norway; suraj.sharma@uit.no (S.S.); hsieh.y.chen@uit.no (Y.-C.H.); jorn.dietze@uit.no (J.D.)

[2] Research Centre for Arctic Petroleum Exploration (ARCEx), Department of Geosciences, UiT The Arctic University of Norway, 9037 Tromsø, Norway; mathias.bockwoldt@uit.no

[3] Department of Biomedicine, University of Bergen, 5020 Bergen, Norway; oyvind.stromland@uib.no (Ø.S.); mathias.ziegler@uib.no (M.Z.)

[4] Department of Clinical Medicine, University of Bergen, 5020 Bergen, Norway

* Correspondence: ines.heiland@uit.no

Abstract: Bacteria use two alternative pathways to synthesize nicotinamide adenine dinucleotide (NAD) from nicotinamide (Nam). A short, two-step route proceeds through nicotinamide mononucleotide (NMN) formation, whereas the other pathway, a four-step route, includes the deamidation of Nam and the reamidation of nicotinic acid adenine dinucleotide (NAAD) to NAD. In addition to having twice as many enzymatic steps, the four-step route appears energetically unfavourable, because the amidation of NAAD includes the cleavage of ATP to AMP. Therefore, it is surprising that this pathway is prevalent not only in bacteria but also in yeast and plants. Here, we demonstrate that the considerably higher chemical stability of the deamidated intermediates, compared with their amidated counterparts, might compensate for the additional energy expenditure, at least at elevated temperatures. Moreover, comprehensive bioinformatics analyses of the available >6000 bacterial genomes indicate that an early selection of one or the other pathway occurred. The mathematical modelling of the NAD pathway dynamics supports this hypothesis, as there appear to be no advantages in having both pathways.

Keywords: NAD biosynthesis; metabolic modelling; kinetic models; phylogenetic analysis

check for updates

Citation: Sharma, S.; Hsieh, Y.-C.; Dietze, J.; Bockwoldt, M.; Strømland, Ø.; Ziegler, M.; Heiland, I. Early Evolutionary Selection of NAD Biosynthesis Pathway in Bacteria. *Metabolites* **2022**, *12*, 569. https://doi.org/10.3390/metabo12070569

Academic Editor: Silvia Sacchi

Received: 13 May 2022
Accepted: 16 June 2022
Published: 21 June 2022

1. Introduction

NAD is an essential cofactor in all organisms. It serves as an electron acceptor for a multitude of redox reactions and is involved in a large number of signalling reactions. In contrast to the reversible interconversion in redox reactions, signalling reactions consume NAD and release nicotinamide (Nam). This needs to be compensated by a constant synthesis of NAD. Two different pathways exist to synthesize NAD from Nam [1,2]. The more common pathway in yeast, plants and bacteria starts with the deamidation of Nam by the bacterial nicotinamidase (PncA) producing nicotinic acid (NA). The Preiss–Handler pathway is employed to convert NA into NAD in a three-step process starting with NA phosphoribosyltransferase (PncB), converting NA into its mononucleotide, NAMN [3]. NAMN is then converted into the dinucleotide NAAD by NA/Nam mononucleotide adenylyltransferase (NadD), which is present in all organisms [4]. The last step is the amidation of NAAD to NAD, catalysed by NAD synthase (NadE). The alternative pathway dominant in higher vertebrates but also present in bacteria [1] converts Nam directly into its mononucleotide, NMN, through nicotinamide phosphoribosyltransferase (Nampt), followed by adenylation through NadD. In addition to synthesis from Nam, NAD can be de novo synthesized from aspartate or tryptophan. An overview of the NAD biosynthesis pathways is shown in Figure 1.

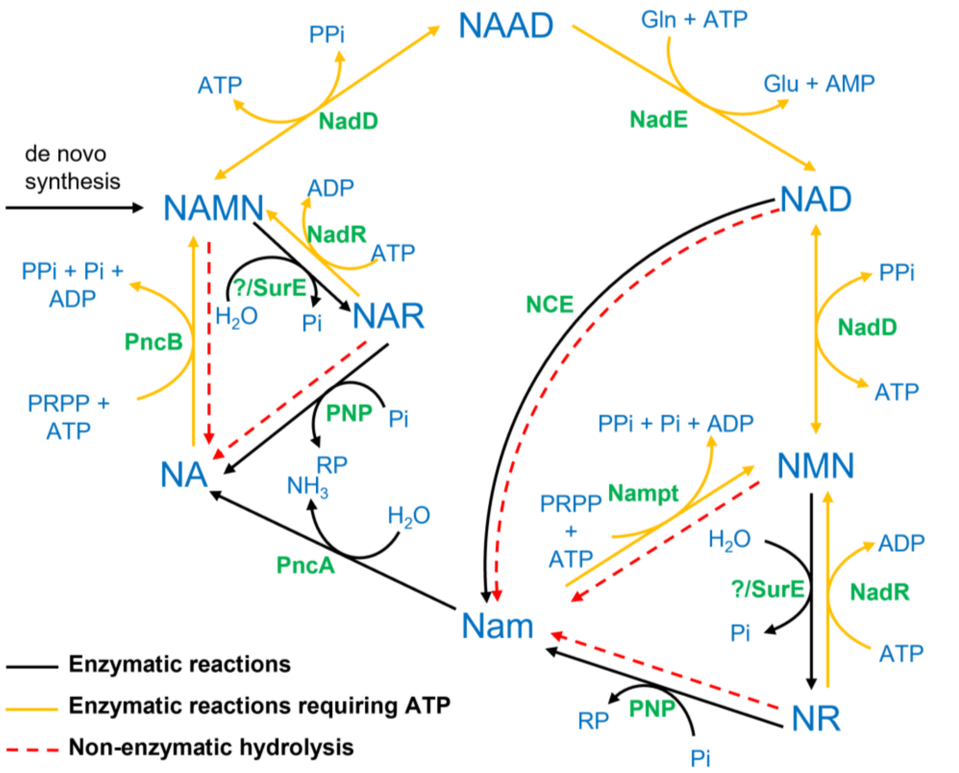

Figure 1. Schematic representation of reactions involved in the biosynthesis of NAD in different organisms. Metabolite names are written in bold blue letters, while the enzyme names are denoted in green colour. The abbreviation NCE (NAD-consuming enzymes) represents all enzymes catalysing signalling reactions consuming NAD. The bold black lines represent enzyme-catalysed reactions, whereas the dotted red lines denote non-enzymatic hydrolysis, mainly occurring at high temperatures. Bold yellow lines are used to denote reactions that require ATP. Whether the bacterial 5′-nucleotidase SurE accepts NMN/NAMN as substrate is unknown; a homologue to the mammalian NT5 could not be identified in bacteria. NAD, Nam adenine dinucleotide; NMN, Nam mononucleotide; NR, Nam riboside; NA, nicotinic acid; NAR, NA riboside; NAMN, NA mononucleotide; NAAD, NA adenine dinucleotide; NadD, NA/NMN adenylyltransferase; Nampt, Nam phosphoribosyltransferase; NadR, NA/Nam riboside kinase; SurE, 5′-nucleotidase; PNP, purine nucleoside phosphorylase; PncA, nicotinamidase; PncB, NA phosphoribosyltransferase; NadE, NAD synthase.

The four-step process that uses PncA requires reamidation to convert NAAD into NAD. This reaction requires ATP releasing AMP [5]; therefore, it is energetically expensive. Thus, the four-step NAD biosynthesis via PncA and NadE requires more ATP than the two-step process that uses Nampt. It is, therefore, surprising that this pathway is dominant in bacteria, yeast and plants. To better understand the evolutionary process that led to preferential selection in bacteria, we performed a detailed phylogenetic analysis of the two pathways focusing on the presence or absence of PncA and Nampt [1,2]. A detailed analysis of this distribution revealed that the selection of the pathways might have occurred based on habitat preferences with a predominant presence of PncA in extremophile and especially in thermophile organisms. We, therefore, measured the thermostability of different pathway intermediates and constructed a temperature-dependent mathematical model including the two alternative pathways of NAD biosynthesis. Model simulations suggest that despite the high thermolysis rates of Nam intermediates, NAD biosynthesis via PncA has no clear advantages over the Nampt pathway at high temperatures. The simulations furthermore indicated that the optimization of NAD biosynthesis indicates that there are no advantages in having both the PncA and the Nampt pathway.

2. Results

2.1. Early Evolutionary Selection of the PncA or the Nampt Pathway in Bacteria

To better understand the prevalence of PncA over Nampt in bacteria, we performed a detailed phylogenetic analysis of the distribution of these enzymes in eukaryotes, bacteria and archaea in more than 8000 organisms (Figure 2A). A previous study of about 200 bacterial genomes described a scattered distribution [2]. In contrast, our phylogenetic analysis of more than 6000 bacteria and archaea detected a mutually exclusive presence of PncA and Nampt, with a clear separation between taxa containing either of the enzymes (Figure 2). This indicates that common ancestors likely harboured both enzymes, with pathway selection having occurred in early evolution.

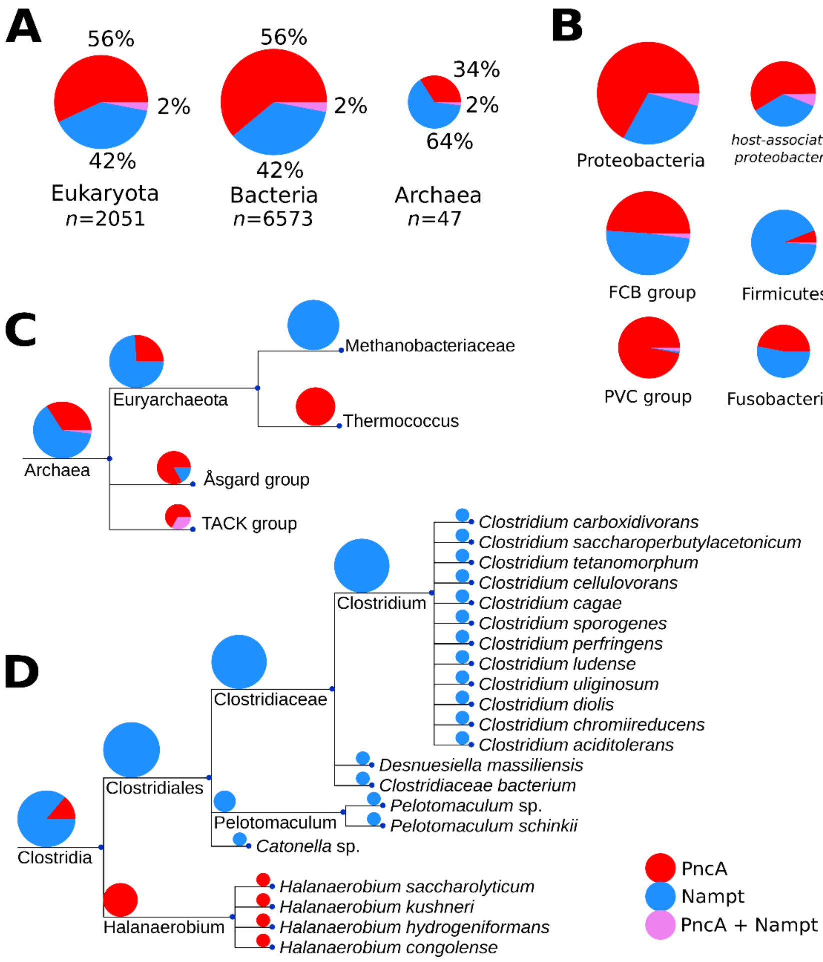

Figure 2. Phylogenetic distribution of PncA and Nampt among the three domains, with expansion into archaea and the bacterial taxa of clostridia. Panel (**A**) shows the total number of organisms (n) represented in the phylogenetic analysis, per domain, as well as their proportions as coloured in the following scheme (PncA, red; Nampt, blue; both PncA and Nampt, violet). The numeric proportion is written beside each segment of the pie chart. The pie chart sizes in each panel are scaled proportionally to the logarithm of the number of organisms represented in each group. In panel (**B**), the enzyme distribution in the largest bacterial clades is shown. Host-association patterns in the proteobacterial group are also shown as an additional pie chart to the right of the total proteobacterial pie chart. Panel (**C**) expands on the domain of archea, particularly to two clades within the phylum euryarchaeota that have presence of only PncA (Thermococcus) or Nampt (Methanobacteriaceae). Panel (**D**) expands on the bacterial class of clostridia, which is a member of the phylum firmicutes. Clostridia is one of several bacterial classes that exhibit a clear preference for PncA in extremophiles, based on our phylogenetic analyses. Branch lengths of the trees are arbitrary.

In archaea, members of genus Thermococcus exclusively harbour PncA, whereas species within the Methanomada group exclusively encode Nampt. We also detected a predominant presence of PncA in thermophilic bacteria. In Clostridia (Figure 2D), for example, the two taxonomic groups show a clear separation between extremophile and non-extremophile organisms. Many of the non-extremophile organisms are pathogenic and/or are known to have vertebrates as hosts. A clear preference is difficult to identify as the habitat information is incomplete, and some organisms can be pathogenic but also live in other habitats such as soil.

Nevertheless, to obtain further insights into a potential role of host association in pathway selection, we used the available habitat information of proteobacteria from the Integrated Microbial Genomes (IMG) database [6,7] and matched organisms found in our phylogenetic analysis with those annotated as being associated with a mammalian host. Annotations for 3154 proteobacterial entries from the IMG were retrieved. Figure 2B shows the host-association patterns of Nampt and PncA within this proteobacterial group. The distribution appears to be similar as in bacteria overall, suggesting that there are no preferences for Nampt in mammalian-host-associated proteobacteria. Thus, host association does not appear to be a dominant selection criterium.

2.2. Higher Glycohydrolysis of Nam Pathway Intermediates at High Temperatures

NAD has been shown to undergo rapid glycohydrolysis at high pH values with Nam and ADP-ribose as products [8,9]. The predominance of PncA in thermophilic organisms may, therefore, be related to the higher chemical stability of acidic NAD biosynthetic intermediates. Therefore, we decided to further analyse the temperature-dependent characteristics of the NAD biosynthesis pathway. As data about the temperature-dependent stability of other intermediates of the NAD biosynthetic pathways were not available, we measured the rate of non-enzymatic hydrolysis of the relevant NAD biosynthesis intermediates at different temperatures in vitro (cf. Table 1). These measurements showed that Nam metabolites such as NMN and NR exhibited higher hydrolysis rates than NAD, especially at high temperatures, whereas the hydrolysis rates of NA intermediates were much lower with no detectable hydrolysis of NAAD (Figure 3) even at 363.15 K (90 °C). These observations suggest a potential advantage of the PncA/Preiss–Handler pathway over the shorter two-step pathway at a higher temperature.

Table 1. Measured hydrolysis rates (%/min) at different temperatures (°C) and calculated pre-exponential factor A and activation energy E_a based on the fitting of the data using the Arrhenius equation (Equation (1)).

Compound	**Temperature (°C)**	**Hydrolysis Rate (%/min)**	**Prefactor** A	**Activation Energy** E_a **(KJ/mol)**
NR	50	0.68 ± 0.04	27083.62	72.41
	75	2.06 ± 0.10		
	90	6.26 ± 0.07		
NMN	50	0.07 ± 0.02	340657.92	82.40
	75	0.62 ± 0.03		
	90	2.76 ± 0.13		
NAD	50	0.00 ± 0.01	1.37×10^{11}	123.16
	75	0.45 ± 0.01		
	90	1.60 ± 0.05		
NAR	50	0.01 ± 0.01	1.25×10^{13}	138.60
	75	0.12 ± 0.02		
	90	0.86 ± 0.02		
NAMN	50	0.00 ± 0.01	2.12×10^{13}	142.49
	75	0.05 ± 0.03		
	90	0.40 ± 0.01		

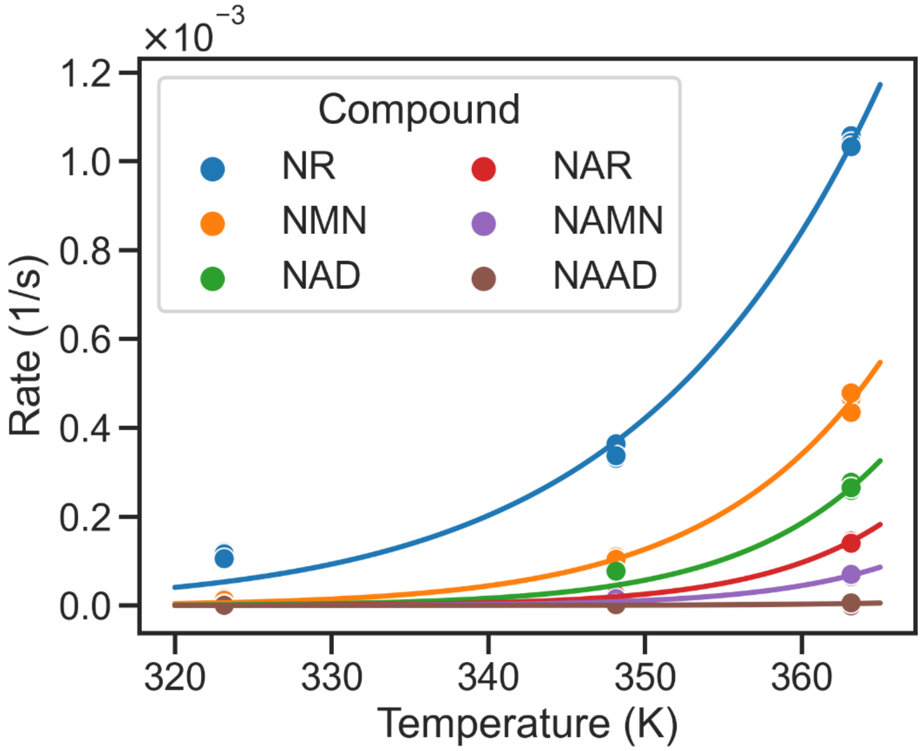

Figure 3. Temperature dependency of non-enzymatic hydrolysis of metabolic intermediates of NAD biosynthesis. Temperature on x-axis is reported in Kelvin (K). Hydrolysis rates are given per second (1/s). While the solid dots represent measured hydrolysis rates, the solid lines denote the temperature-dependent hydrolysis rates calculated using a fitted Arrhenius equation (cf. Table 1). NAAD shows the lowest hydrolysis rates, while NR exhibits the highest glycohydrolysis rate at 363.15 K.

2.3. PncA Pathway Is More Energy Demanding Even at High Temperatures

To analyse the potential effects of the non-enzymatic hydrolysis at high temperatures of nicotinamide intermediate on the efficiency of NAD biosynthesis, we developed a mathematical model of NAD biosynthesis using temperature-dependent rate laws. For the purpose, we integrated relevant enzymatic reactions and rate laws from our previously published NAD model (BioModels: MODEL1905220001; [1]). Additionally, to make the model temperature dependent, the reaction rates were scaled using the Arrhenius equation (cf. Equation (1)). As no measured activation energies were available for the enzymatic reactions, we set the activation energies such that the reactions were thermodynamically feasible at physiological conditions and had a Q10 within the range of 2–3, as previously conducted [10,11]. Further, we added non-enzymatic hydrolysis reactions using rates and activation energies determined by fitting the experimentally measured data (see Table 1) to the Arrhenius equation. The resulting model is based on a set of ordinary differential equations (ODEs) that can simulate temperature-dependent steady-state changes in the pathway intermediates of NAD biosynthesis via PncA and Nampt (for details see Supplementary Equations (S1)–(S8)).

To estimate the energy efficiency of the two alternative biosynthetic routes, we calculated the ratio between ATP consumption and NAD production. ATP consumption is defined as the total flux of ATP-consuming reactions (J_{ATP}) and is caused by the following reactions catalysed by: Nampt, PncB, NadE, NadD and NadR. The amidation of NAAD uses one ATP molecule to produce NAD and one AMP. As two phosphorylation steps are required to convert AMP back to ATP, we represent the energy demand of the NadE catalysed reaction by multiplying its flux by 2. The total flux of ATP-consuming reactions is thus:

$$J_{ATP} = J_{NAMPT} + J_{PNCB} + 2 \cdot J_{NADE} + J_{NADD} + J_{NADR}. \tag{1}$$

The synthesis of the mononucleotides through Nampt or PncB is non-stoichiometrically coupled to the hydrolysis of ATP owing to the autophosphorylation of a histidine residue, thereby increasing substrate affinity. For NAMPT, it has been estimated that about one ATP is consumed per catalytic cycle at an ATP concentration of 2–2.5 mM [12]. Given the highly similar reaction and enzyme structures, this ATP requirement can be assumed to be the same for both enzymes; therefore, it has most likely no or little influence on the efficiency of either pathway.

NAD production is attributed to the flux generated by NadE (J_{NADE}) in the PncA pathway and NadD (J_{NADD}) in the Nampt pathway converting NAAD and NMN to NAD, respectively:

$$J_{NAD} = J_{NADE} + J_{NADD} \tag{2}$$

To investigate the efficiency of the two biosynthetic routes of NAD production, we assumed that the bacterial NAD biosynthetic pathway adjusts the concentration of metabolic enzymes to minimize the energy demand defined by the ratio of the total flux of ATP-consuming reactions (J_{ATP}) to the NAD production flux (J_{NAD}) given a range of free-NAD concentration. The optimization problem can, therefore, be represented as:

$$\begin{gathered} \underset{E\in[1^{-10},100]}{min} \frac{J_{ATP}}{J_{NAD}} \\ subject\ to:\ 0.1 < C_{NAD} < 0.3 \\ B = \{x \vee 0.01 \leq x \leq 100\} \\ \sum_i E \leq 1000; i = 1, \ldots, n \end{gathered} \tag{3}$$

This represents the value of the argument E ($E = A \cup B$, where $A = \{E_{PNCA}, E_{NAMPT}\}$ and $B = \{E_{PNCB}, E_{NADD}, E_{NADR}, E_{PNP}, E_{NADE}, E_{NCE}, E_{SURE}\}$) in the interval [$1^{-10}$, 100] nM that minimizes objective function J_{ATP}/J_{NAD}, with the added constraints that the steady-state concentration of NAD is between 0.1 and 0.3 mM; the sum of concentrations of n metabolic enzymes is less than 1000 nM; and the subset B is the set of all elements x, such that x is in the interval [0.01, 100] nM. The first inequality constraint ensures the experimentally measured free-NAD concentrations in bacteria, whereas the latter two constrain the lower and upper limits on the chosen concentrations of metabolic enzymes. Optimization was performed using the Evolutionary Programming algorithm of COPASI 4.29 [13], with the number of generations and population size being set to 200 and 20, respectively. The selection of these optimization parameters was based on the convergence of the desired output within the tolerance range of 1×10^{-6}.

For the given optimisation problem, we found that multiple combinations of enzyme abundances could yield the desired steady-state concentration of NAD while achieving the lowest possible J_{ATP}/J_{NAD} ratio. Therefore, we repeated the optimization 1000 times and analysed the distribution of enzyme abundances. The optimisation results show that the optimum was found when the amount of either PncA or Nampt was very small in comparison to the other enzymes (cf. Figure 4A). This observation indicates that under optimal conditions, only one of the two pathways is active. Consequently, it appears reasonable that bacteria typically only have either PncA or Nampt but rarely both. We used this behaviour to classify the results of the modelling into subsets with different enzyme abundances, where *Nampt << PncA* reflects model solutions optimised for the PncA pathway and vice versa. This enabled us to analyse the distribution of the overall enzyme abundances that facilitate NAD biosynthesis via the PncA and Nampt pathways (Figure 4B,C). Based on the simulation results, both J_{ATP} and J_{NAD} increased with the increase in temperature (Figure 4D,E). The ratio of ATP consumption to NAD production of the PncA pathway was always higher than that of the Nampt pathway (see Figure 4F). We used metabolic control analysis [14,15] to investigate the control exerted by metabolic enzymes on the steady-state concentrations of pathway intermediates at different temperatures. PncA had a high positive control coefficient on NAD concentration and a strong negative control coefficient on Nam concentrations at all temperatures, whereas Nampt had a positive concentration control coefficient for both Nam and NAD. See Supplementary Figures S1 and S2 for details. The steady-state concentrations of the metabolic intermediates of NAD biosynthesis via the PncA and Nampt pathways are shown in Supplementary Figure S3.

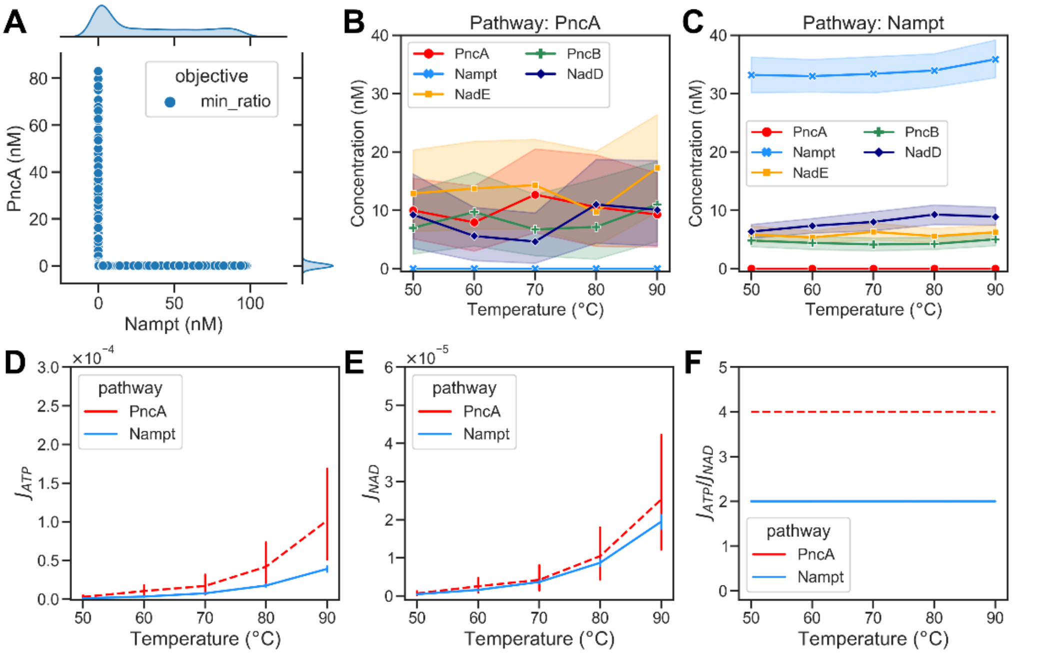

Figure 4. Optimization results of minimizing the energy demands of bacterial NAD biosynthetic pathways. (**A**) A scatter plot showing the optimized abundance of Nampt versus PncA when minimizing the ratio of ATP consumption to NAD production required to maintain the required NAD steady-state concentration (0.1 mM < C_{NAD} < 0.3 mM). Each dot represents an optimisation result. The marginal plots show kernel density estimate plots that show the enzyme abundance distributions in the optimisation results. (**B**) Abundance of metabolic enzymes at different temperatures in the PncA pathway. (**C**) Optimized abundance of metabolic enzymes in Nampt pathway. (**D**) The ATP consumption flux, J_{ATP}, in the PncA pathway (dashed red curve) compared with the Nampt pathway (blue curve) at different temperatures. (**E**) The NAD production flux, J_{NAD}, in the PncA pathway compared with that in the Nampt pathway at different temperatures. (**F**) The ratio of ATP consumption to NAD production at different temperatures in PncA and Nampt pathways.

2.4. *Optimal Pathway Performance Achieved with Mutually Exclusive Presence of PncA and Nampt*

As we do not know what the bacterial NAD biosynthesis is optimized for in nature, we analysed two alternative objectives: (1) minimization of ATP consumption and (2) maximization of NAD concentration. For both scenarios, we optimized the abundance of the metabolic enzymes of NAD biosynthesis for the desired objective. Optimization was performed as described earlier. To minimize ATP consumption, we formulated the optimization problem as follows:

$$\begin{gathered}\underset{E\in[1^{-10},100]}{min}\ J_{ATP}\\ subject\ to:\ 0.1 < C_{NAD} < 0.3\\ B = \{x \vee 0.01 \leq x \leq 100\}\\ \sum_i E \leq 1000; i = 1,\ldots,n\end{gathered} \tag{4}$$

This represents the value of the argument E ($E = A \cup B$, where $A = \{E_{PNCA}, E_{NAMPT}\}$ and $B = \{E_{PNCB}, E_{NADD}, E_{NADR}, E_{PNP}, E_{NADE}, E_{NCE}, E_{SURE}\}$) in the interval [$1^{-10}$, 100] nM that minimizes objective function J_{ATP}, with the added constraints that the steady-state concentration of NAD is between 0.1 and 0.3 mM,; the subset B is the set of all elements x, such that x is in the interval [0.01, 100] nM; and the sum of concentrations of n metabolic enzymes is less than 1000 nM. As with the previous optimization, we found that several different combinations of enzyme abundances could yield the same optimal output, and most solutions had low abundance of either the PncA or Nampt enzyme. In the PncA pathway (Nampt << PncA), NadE is the enzyme that directly affects NAD production. Thus, a higher concentration of NadE with the increase in temperature was observed, compensating the rapid hydrolysis of NAD at high temperatures. Additionally, increased abundance of PncB appeared to be important to efficiently replenish NAD. It should be noted that the substrate of PncB is NA, which, in turn, is the product of the non-enzymatic hydrolysis reactions. Thus, increasing the abundance of PncB helped to channel the flux towards the production of NAD (Figure 5B). In addition, a high abundance of NadD appeared to be advantageous. The latter was also observed for the Nampt pathway. A further analysis of the pathway fluxes showed that despite the optimal distribution of metabolic enzymes in the PncA pathway, the NAD metabolism did still require more ATP than the Nampt pathway even at high temperatures (cf. Figure 5D–F). The model-simulated steady-state concentrations of the metabolic intermediates of NAD biosynthesis via the PncA and Nampt pathways are shown in Supplementary Figure S3.

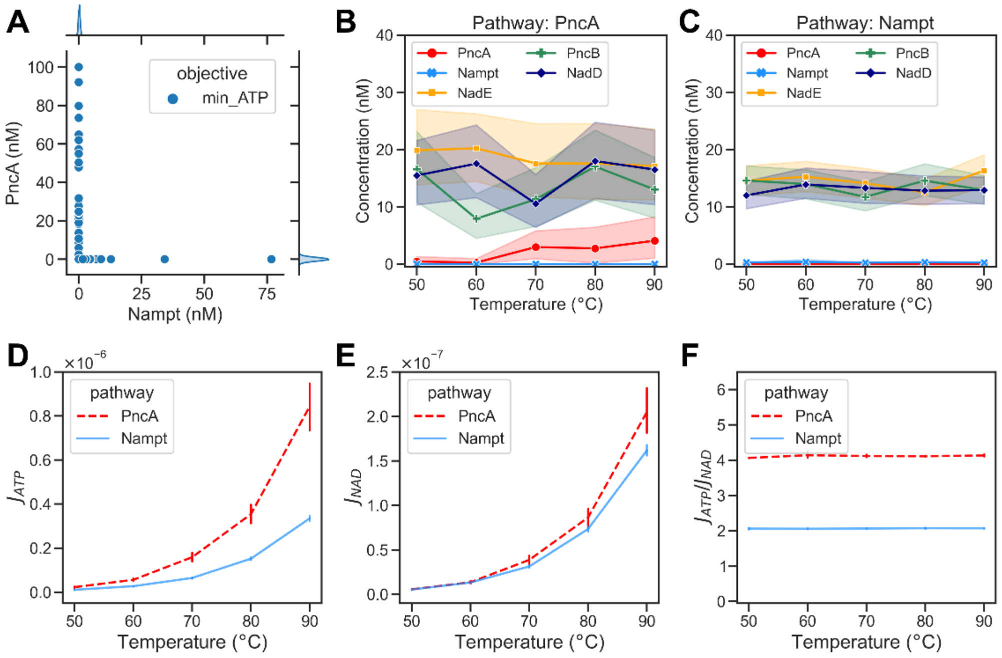

Figure 5. Optimization results of minimizing the total flux of ATP-consuming reactions of bacterial NAD biosynthetic pathways. (**A**) A scatter plot showing the abundances of Nampt versus PncA when minimizing the ATP consumption required to maintain the required NAD steady-state concentration (0.1 mM < C_{NAD} < 0.3 mM). Each dot represents an optimisation result. The marginal plots show kernel density estimate plots that show the enzyme abundance distributions in the optimisation results. (**B**) Optimized abundance of enzymes at different temperatures in PncA pathway. (**C**) Optimized abundance of enzymes from Nampt pathway. (**D**) The ATP consumption flux, J_{ATP}, in PncA (dashed red curve) and Nampt (blue curve) pathways at different temperatures. (**E**) Computed NAD production flux, J_{NAD}, in PncA and Nampt pathways at different temperatures. (**F**) Calculated ratio of ATP consumption flux to NAD production flux at different temperatures in PncA and Nampt pathways.

We then maximized the NAD concentration in our models, using the following objective function:

$$\begin{gathered}\underset{E\in[1^{-10},100]}{max} \quad C_{NAD}\\ subject\ to:\ B=\{x\vee 0.01\le x\le 100\}\\ \sum_i E\le 1000; i=1,\ldots,n\end{gathered} \tag{5}$$

where the argument E ($E = A\cup B$, where $A=\{E_{PNCA},E_{NAMPT}\}$ and $B=\{E_{PNCB},E_{NADD},E_{NADR},E_{PNP},E_{NADE},E_{NCE},E_{SURE}\}$) in the interval [$1^{-10}$, 100] nM maximizes the NAD concentration (C_{NAD}), with the added constraints that the sum of concentrations of n metabolic enzymes is less than 1000 nM and the subset B is the set of all elements x, such that x is in the interval [0.01, 100] nM. We again predominantly found solutions with low abundance of either PncA or Nampt. The PncA pathway exhibited an optimal enzyme combination with a high abundance of NadE (cf. Figure 6A). This can be understood intuitively, as the system tries to compensate for the loss of NAD through high glycohydrolysis, especially at high temperatures. The ATP consumption flux in the PncA pathway was higher than that in the Nampt pathway at different temperatures, while the NAD production flux in both pathways was similar (Figure 6C,D). Consequently, the PncA pathway was found to consume more ATP per NAD produced at all temperatures (Figure 6E).

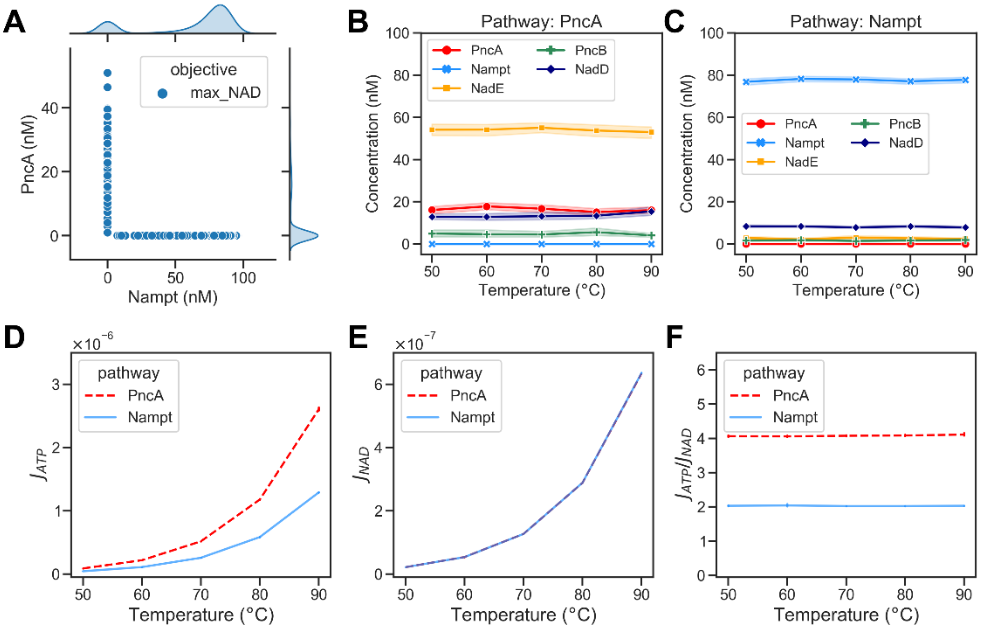

Figure 6. Optimization results of maximizing the NAD concentration of bacterial NAD biosynthetic pathways. (**A**) A scatter plot showing the abundances of Nampt versus PncA when maximizing the NAD production ratio. Each dot represents an optimisation result. The marginal plots show kernel density estimate plots that show the enzyme abundance distributions in the optimisation results. (**B**) Optimized abundance of enzymes at different temperatures in PncA pathway. (**C**) Optimized abundance of enzymes from Nampt pathway. (**D**) The ATP consumption flux, J_{ATP}, in PncA (dashed red curve) and Nampt (blue curve) pathway at different temperature (**E**) The NAD production flux, J_{NAD}, in PncA and Nampt pathways at different temperatures. (**F**) The ratio of ATP consumption flux to NAD production flux at different temperatures in PncA and Nampt pathways.

3. Discussion

The enigmatic, scattered distribution of different NAD pathways in bacteria has been a matter of several previous investigations. With more than 6000 bacterial genomes available to date, we are now able to obtain a more detailed insight into the pathway distribution and its potential evolution. Our comprehensive phylogenetic analysis suggests an early evolutionary selection of either the PncA or the Nampt pathway with the respective common ancestors having both enzymes. As the PncA pathway is energetically less efficient, it is surprising that we found it in more than 50% of the analysed bacteria. We, therefore, looked at a potential role of bacterial habitats in selection preferences. This is, however, difficult, as (1) habitat information is limited and (2) habitats can change, especially over the long evolutionary timeframe we investigated. We did, however, see a predominance of the PncA pathway in extremophile organisms, indicating a potential advantage under these conditions. The measurement of the chemical stability of NAD pathway intermediates at high temperatures supported this hypothesis, as the nicotinic acid intermediates of the PncA pathway were much more stable than the nicotinamide intermediates of the Nampt pathway.

Mathematical modelling approaches allowed us to analyse hypothetical scenarios such as the dynamics of the two alternative pathways at different temperatures. We, therefore, used our ODE-based kinetic model of NAD biosynthesis to simulate temperature-dependent changes in enzyme-catalysed reaction rates. We, unfortunately, do not know what organisms are optimized for in nature; we, therefore, analysed three different objectives, i.e., (1) minimizing the ratio of ATP consumption to NAD production, (2) minimizing overall ATP consumption and (3) maximizing free-NAD concentrations. Interestingly, we found that independently of the objective function, optimization did predominantly result in models having either the PncA pathway or the Nampt pathway but rarely both. This indicates that there are no advantages in having both pathways, supporting both the early evolutionary selection and the disappearance of early ancestors that had both pathways. However, despite the higher glycohydrolysis rate of Nampt pathway intermediates, our modelling approach was not able to find any energetical advantage of the PncA pathway, as the ATP consumption was higher than that of the Nampt pathway at all temperatures. There are, of course, many parameters that we did not consider in our modelling approach, such as the thermostability of the enzymes, the achievable temperature optima and the exact activation energy for the different enzymatic reactions. Analysing the contribution of these parameters is difficult as very limited experimental data are available, leading to an unconstrained model that enables all possible scenarios (not shown). It is, furthermore, important to note that we forced the model to maintain a certain NAD concentration, while it has been shown that the high hydrolysis rates of NAD at high temperatures can impose a problem for thermophiles and hyperthermophiles [16]. Given the high stability of NAAD, it appears that a likely strategy could be to rather produce NAD on demand and thus regulate NAD production through the regulation of NadE. Unfortunately, there are very limited expression data available for thermophilic organisms at different temperatures that could support this hypothesis. We did not consider potential contributions by the de novo synthesis from aspartate either, which could compensate for the high hydrolysis of pathway intermediates. Thus, further investigations are required to better understand how extremophile organisms maintain NAD synthesis despite the high non-enzymatic hydrolysis of NAD and its intermediates.

4. Materials and Methods

4.1. Phylogenetic Analysis of Enzyme Distributions

We constructed a comprehensive overview of the taxonomic distributions of our target enzymes by performing query searches of our enzymes against the National Center for Biotechnology Information (NCBI) non-redundant protein database (nr) and subsequently combining and remapping them to the general taxonomic tree as constructed in NCBI. More specifically, we used functionally verified protein sequences of PncA and Nampt from several model organisms (see Supplementary Table S1), using their protein sequences as queries via Blastp [17]. The total number of hits returned per query was limited to 5000 sequences, and the alignment parameters followed the defaults: BLOSUM62 substitution matrix; word size, 6; gap open penalty, 11; and gap extension, 1, which was used to ensure stringent sequence selection and to avoid the inclusion of functional related but non-homologous sequences. We additionally filtered each set of resultant sequences based on cut-offs for protein length and e-value to optimize for sequence coverage and hit significance. The protein length cut-off was determined by a consensus of the hit length as taken from a histogram of all hit lengths, and the e-value cut-off was chosen as the lowest e-value at which there were sequence cross hits between our various resultant sequence sets. The cut-off values per enzyme are given in the Supplementary Table S1. The aim of sequence filtering through these cut-offs was to eliminate false positive hits across our query protein set. Taxonomic information per sequence was taken directly from the Blast results (XML2 format, taxid field). This methodology has been used previously [1].

4.2. Thermostability of NAD and Related Metabolites Determined by ^{1}H-NMR

The compounds were dissolved in NMR buffer (25 mM sodium phosphate, pH 5.8, and 5% (v/v) D_2O) and diluted to a final concentration of 500 µM. The samples were incubated at the indicated temperature for 10 min and chilled on ice for 5 min before measurement. NMR data were collected on a Bruker Ascend 850 MHz instrument fitted with a cryogenically cooled triple-resonance 5 mm TCI probe with pulsed-field gradients along the z-axis at 23 °C. The samples were measured by ^{1}H-NMR using the pulse sequence zgesgppe, allowing water suppression to be achieved using excitation sculpting with pulsed-field gradients and perfect echo. The spectra were acquired with 64 scans and a recovery delay of 3.9 s. The spectra were assigned using standard correlation methods. Resonances of interest were integrated using the program Dynamics Center 2.5 (Bruker; Bremen Germany) and compared to baseline spectra.

4.3. Rate of Temperature-Dependent Non-Enzymatic Hydrolysis

Based on experimentally measured non-enzymatic hydrolysis rates (see Table 1), pre-exponential factor A and activation energy E_a were inferred by fitting the data to the Arrhenius relation [18]:

$$k = Ae^{\frac{E_a}{RT}}. \tag{6}$$

Using the values for A and E_a, rate constant k could be calculated in dependence of absolute temperature T [10].

4.4. Mathematical Modelling of NAD Metabolism in Bacteria

For the simulation of NAD pathway dynamics, we consider a dynamic system C of n variables, which is defined as:

$$\frac{dC}{dt} = f(C,k), \tag{7}$$

where reaction rates are denoted by parameter k. The temperature dependence of the rate laws is modelled using the Arrhenius equation (Equation (6)) as described earlier [10]. The complete list of ordinary differential equations that constitute the dynamic system of model variables are given in Supplementary Materials (cf. Equations (S1)–(S8)). The model described here uses a subset of reactions (see Figure 1), and the corresponding kinetic constants are taken from enzyme database BRENDA [19] and additionally evaluated by checking the original literature for comparable measurement conditions. Parameter values were used as in our previously published model of NAD metabolism (BioModels: MODEL1905220001; [1]). For details, see supplementary text. The initial concentration of free NAD was set to 0.3 mM, corresponding to experimentally measured free-NAD concentrations. The initial concentration of all other pathway intermediates was set to 0, unless stated otherwise. The model did converge on the same steady state as long as the sum of the concentration of all Nam-containing intermediates remained 0.3 mM. Influx and growth were neglected in the simulations to reduce the complexity and number of parameters. All parameters were assumed to hold for base temperature T_0 = 37.5 °C.

4.5. Software and Data

The temperature-dependent model of NAD metabolism was submitted to the BioModels database (model MODEL2103290001). The ETE3 toolkit v3.1.2 [20] was used to visualize the phylogenetic trees. The python scripts used to perform the phylogenetic analysis are available on GitHub: https://github.com/MolecularBioinformatics/Phylogenetic-analysis (accessed on 25 March 2019). The steady-state calculation of fluxes and concentrations was performed using COPASI 4.29 [11]. The Python scripts used to calculate the modelling results of this paper can be downloaded from GitHub: https://github.com/MolecularBioinformatics/thermophilesNAD (accessed on 12 May 2022).

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/metabo12070569/s1, Figure S1: Concentration control coefficients of NAD biosynthesis with exclusive presence of PncA enzyme. Each subplot shows the concentration control coefficient on the metabolic intermediates of PncA pathway due to the perturbation in the enzyme concentration corresponding to an enzyme catalyzed reaction (y-axis) at different temperatures (x-axis). Refer the computational model submitted at the BioModels database as MODEL2103290001 for details about the reactions; Figure S2: Concentration control coefficients of NAD biosynthesis via exclusive presence of Nampt enzyme. Each subplot shows the concentration control coefficient on the metabolic intermediates of Nampt pathway due to the perturbation in the enzyme concentration corresponding to an enzyme catalyzed reaction (y-axis) at different temperatures (x-axis). Refer the computational model submitted at the BioModels database as MODEL2103290001 for details about the reactions; Figure S3: Simulated steady-state concentrations of metabolic intermediates of NAD biosynthesis via PNCA and NAMPT pathway at different temperatures using the mathematical model of NAD biosynthesis. Model-simulated concentrations of pathway intermediates while A) minimization of the ratio of ATP consumption (J_{ATP}) to NAD production flux (J_{NAD}), B) minimization of the ATP consumption flux (J_{ATP}), and C) maximization of the free NAD concentration (C_{NAD}). Error bars show the different simulated concentrations due to different combinations of enzyme concentrations found as a solution to the optimization problem at a given temperatures; Table S1: Seed sequences for phylogenetic analysis: enzyme identity and cutoffs for minimum length and maximum E-value; Table S2: An overview of rate laws used by the metabolic enzymes of NAD pathway. Ref. [21] is cited in the Supplementary Materials file.

Author Contributions: S.S. and J.D. built the mathematical model and performed the simulations. Y.-C.H. and M.B. performed the phylogenetic analyses. Ø.S. performed the experimental measurements. M.Z. and I.H. designed the study. All authors were involved in the preparation and revision of the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: M.Z., Ø.S., J.D. and I.H. were supported by grants from Norwegian Research Council (302314, 315849 and 325172).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The model file can be downloaded from BioModels: MODEL2103290001. The Python scripts used to calculate the results shown in this paper can be downloaded from Github: https://github.com/MolecularBioinformatics/thermophilesNAD (accessed on 25 March 2022) and https://github.com/MolecularBioinformatics/Phylogenetic-analysis (accessed on 10 July 2019).

Acknowledgments: We made use of the Facility for Biophysics, Structural Biology and Screening at the University of Bergen (BiSS), which has received funding from the RCN through the NORCRYST (grant number 245828) consortium. This work was also supported by the RCN through the Norwegian NMR Platform, NNP (226244/F50). The APC for this article has been covered by UiT The Arctic University of Norway.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Bockwoldt, M.; Houry, D.; Niere, M.; Gossmann, T.I.; Reinartz, I.; Schug, A.; Ziegler, M.; Heiland, I. Identification of evolutionary and kinetic drivers of NAD-dependent signaling. *Proc. Natl. Acad. Sci. USA* **2019**, *116*, 15957–15966. [CrossRef] [PubMed]
2. Gazzaniga, F.; Stebbins, R.; Chang, S.Z.; McPeek, M.A.; Brenner, C. Microbial NAD Metabolism: Lessons from Comparative Genomics. *Microbiol. Mol. Biol. Rev.* **2009**, *73*, 529–541. [CrossRef] [PubMed]
3. Preiss, J.; Handler, P. Biosynthesis of diphosphopyridine nucleotide. I. Identification of intermediates. *J. Biol. Chem.* **1958**, *233*, 488–492. Available online: http://www.jbc.org/content/233/2/488.short (accessed on 12 May 2022). [CrossRef]
4. De Figueiredo, L.F.; Gossmann, T.I.; Ziegler, M.; Schuster, S. Pathway analysis of NAD + metabolism. *Biochem. J.* **2011**, *439*, 341–348. [CrossRef] [PubMed]
5. Spencer, R.L.; Preiss, J. Biosynthesis of diphosphopyridine nucleotide. The purification and the properties of diphospyridine nucleotide synthetase from *Escherichia coli* b. *J. Biol. Chem.* **1967**, *242*, 385–392. Available online: http://www.ncbi.nlm.nih.gov/pubmed/4290215 (accessed on 12 May 2022). [CrossRef]
6. Mukherjee, S.; Stamatis, D.; Bertsch, J.; Ovchinnikova, G.; Sundaramurthi, J.C.; Lee, J.; Kandimalla, M.; Chen, I.-M.A.; Kyrpides, N.C.; Reddy, T.B.K. Genomes OnLine Database (GOLD) v.8: Overview and updates. *Nucleic Acids Res.* **2021**, *49*, D723–D733. [CrossRef] [PubMed]
7. Chen, I.-M.A.; Chu, K.; Palaniappan, K.; Ratner, A.; Huang, J.; Huntemann, M.; Hajek, P.; Ritter, S.; Varghese, N.; Seshadri, R.; et al. The IMG/M data management and analysis system v.6.0: New tools and advanced capabilities. *Nucleic Acids Res.* **2021**, *49*, D751–D763. [CrossRef] [PubMed]
8. Kaplan, N.O.; Colowick, S.P.; Barnes, C.C. Effect of alkali on diphosphopyridine nucleotide. *J. Biol. Chem.* **1951**, *191*, 461–472. Available online: http://www.ncbi.nlm.nih.gov/pubmed/14861192 (accessed on 12 May 2022). [CrossRef]
9. Colowick, S.P.; Kaplan, N.O.; Ciotti, M.M. The reaction of pyridine nucleotide with cyanide and its analytical use. *J. Biol. Chem.* **1951**, *191*, 447–459. Available online: http://www.ncbi.nlm.nih.gov/pubmed/14861191 (accessed on 12 May 2022). [CrossRef]
10. Bodenstein, C.; Heiland, I.; Schuster, S. Calculating activation energies for temperature compensation in circadian rhythms. *Phys. Biol.* **2011**, *8*, 056007. [CrossRef] [PubMed]
11. Heiland, I.; Bodenstein, C.; Hinze, T.; Weisheit, O.; Ebenhoeh, O.; Mittag, M.; Schuster, S. Modeling temperature entrainment of circadian clocks using the Arrhenius equation and a reconstructed model from Chlamydomonas reinhardtii. *J. Biol. Phys.* **2012**, *38*, 449–464. [CrossRef] [PubMed]
12. Burgos, E.S.; Schramm, V.L. Weak Coupling of ATP Hydrolysis to the Chemical Equilibrium of Human Nicotinamide Phosphoribosyltransferase. *Biochemistry* **2008**, *46*, 11086–11096. [CrossRef] [PubMed]
13. Hoops, S.; Sahle, S.; Gauges, R.; Lee, C.; Pahle, J.; Simus, N.; Singhal, M.; Xu, L.; Mendes, P.; Kummer, U. COPASI—A COmplex PAthway SImulator. *Bioinformatics* **2006**, *22*, 3067–3074. [CrossRef] [PubMed]
14. Heinrich, R.; Rapoport, T.A. A Linear Steady-State Treatment of Enzymatic Chains. *Eur. J. Biochem.* **1974**, *42*, 89–95. [CrossRef] [PubMed]
15. Heinrich, R.; Schuster, S. *The Regulation of Cellular Systems*; Springer US: Boston, MA, USA, 1996.
16. Hachisuka, S.-I.; Sato, T.; Atomi, H. Metabolism Dealing with Thermal Degradation of NAD+ in the Hyperthermophilic Archaeon *Thermococcus kodakarensis*. *J. Bacteriol.* **2017**, *199*, e00162-17. [CrossRef] [PubMed]
17. Altschul, S.F.; Gish, W.; Miller, W.; Myers, E.W.; Lipman, D.J. Basic local alignment search tool. *J. Mol. Biol.* **1990**, *215*, 403–410. [CrossRef]
18. Arrhenius, S. Über die Dissociationswärme und den Einfluss der Temperatur auf den Dissociationsgrad der Elektrolyte. *Z. Phys. Chem.* **1889**, *4*, 96–116. [CrossRef]
19. Schomburg, I.; Chang, A.; Schomburg, D. BRENDA, enzyme data and metabolic information. *Nucleic Acids Res.* **2002**, *30*, 47–49. [CrossRef] [PubMed]
20. Huerta-Cepas, J.; Serra, F.; Bork, P. ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data. *Mol. Biol. Evol.* **2016**, *33*, 1635–1638. [CrossRef] [PubMed]
21. Kacser, H.; Burns, J.A. The control of flux. *Symp. Soc. Exp. Biol.* **1973**, *27*, 65–104. Available online: http://www.ncbi.nlm.nih.gov/pubmed/4148886 (accessed on 17 May 2017).