

Beyond sequence similarity: Computational classification of plant enzymes through electrostatic similarity

Matthew Donald Mortimer

24 December 2025

A thesis submitted for the degree of
Doctor of Philosophy
of
The Australian National University

Research School of Biology
The Australian National University

Declaration of Originality

I, Matthew Donald Mortimer, declare that:

1. This thesis is my own original work, except where otherwise acknowledged.
2. The thesis has not been submitted for a degree at any other university.
3. The research presented in this thesis was conducted under the supervision of Dr Andrew F. Bowerman, Prof. Barry J. Pogson.
4. Where the research was conducted jointly with others, I have clearly indicated the nature and extent of my contribution.
5. Claude AI was used as an editing service to proof read, aid in clarity and improve flow. All writing, content, analysis, interpretation and conclusions remain my own, except where otherwise acknowledged. Claude AI was used for code refactoring, debugging and to aid in code documentation.
6. The thesis contains approximately 59,100 words.

Signature

Matthew Donald Mortimer

Date

Acknowledgements

I would like to express my deepest gratitude to all those who have supported and assisted me throughout the course of my PhD research and writing. This thesis would not have been possible without the contributions, encouragement, and generosity of many individuals and institutions.

Supervisors

First and foremost, I am profoundly grateful to my supervisors, **Professor Barry Pogson** and **Dr Andrew Bowerman**, for their invaluable guidance, expertise, and unwavering support throughout this journey. Their intellectual insight, patience, and encouragement have shaped not only this thesis but also my growth as a researcher. I thank Barry for his support throughout my PhD including the opportunities and resources he provided. Andrew's guidance, friendship and support has been invaluable, without his insight, empathy discussions and ideas this thesis would not have been written.

I would also like to thank my co-supervisors — **Associate Professor Caitlin Byrt**, **Associate Professor Simon Williams**, and **Professor Colin Jackson** — for their constructive feedback, helpful discussions, and ongoing support across different stages of the research. Their diverse perspectives and encouragement were instrumental in strengthening the direction and quality of this work.

I would like to extend particular thanks to **Dr Frank Röber** who has been extremely generous with his time, providing me invaluable guidance and support as my mentor throughout my PhD. Thanks to **Dr Diep Ganguly** who's friendship, regular chats as well as advice were immensely valuable and always enjoyable.

Technical & Administrative Support

I am indebted to **Dr Derek Collinge**, whose skill, organisation, and attention to detail kept the laboratory running smoothly. Their support in managing equipment, resources, and logistics made much of this research possible.

I would also like to acknowledge the exceptional efforts of the ANU RSB IT Services team, particularly **Sharyn Wragg**, who provided crucial technical support when-ever it was needed. Their assistance with software, computing resources, and trouble-shooting was indispensable to the progress of this project. In particular their support in setting up a stand-alone linux workstation made this work possible. Without this and the permissions to make deep changes to the system this work would not have been possible.

Colleagues & Research Community

My heartfelt thanks go to all the members of the Pogson Group, past and present, who contributed to a collaborative and stimulating research environment. The discussions, shared expertise, and camaraderie in the lab greatly enriched both the work and the experience of doing a PhD.

I am also grateful to colleagues in the wider Research School of Biology, whose feedback during seminars, workshops, and informal conversations helped refine many of the ideas presented here.

Family & Personal Support

I would like to thank my parents for their continued support and interest. Most importantly, I would like to thank my wife, Ira, for her love, patience, and unwavering belief in me. This PhD has been a significant burden on her as well, which she has not complained about one. Her support during the highs and lows of this PhD sustained me more than words can express.

I also extend my deepest gratitude to my family and close friends, whose encouragement and understanding gave me the strength to persevere.

Abstract

Thesis significance statements

- **Methodological Approach:** This thesis applies a progression of computational classification methods to plant enzymes (phytoene synthase, [DXO1](#), [eIF4E](#), and the stress-sensing nucleotide phosphatases). Beginning with [SSNs](#) and [Evolocity](#) network analysis for large-scale classification, progressing through molecular phylogenetics for evolutionary inference, then structural similarity comparison, and culminating in a novel Fourier-decomposed electrostatic field comparison approach. Each method revealed limitations that motivated the next, with the electrostatic approach detecting functional relationships invisible to sequence and structural methods.
- **Key Evolutionary Discovery:** Functional relationships can diverge from phylogenetic relationships. I show phylogenetically distinct proteins can share catalytic properties (Algal [AHL](#) and Ancient [CNP](#) clustering electrostatically with SAL), while phylogenetically similar proteins can be functionally divergent ([TaSAL1](#)/[TaSAL2](#) with opposing yield effects; [AHL- \$\alpha\$](#) /[AHL- \$\beta\$](#) with differential pH sensitivity).
- **Biological Significance:** Neo-functionalisation following whole-genome duplication is pervasive across plant protein families, with novel regulatory functions arising in [DXO1](#), [eIF\(ISO\)4E](#), and [AHL](#) sub-clades. The electrostatic clustering of Algal [AHL](#) and Ancient [CNP](#) with SAL suggests continuous [PAP](#) catabolic capacity across Viridiplantae, resolving an apparent functional gap spanning hundreds of millions of years where lineages lacking SAL appeared to have no enzyme for this essential function. This approach demonstrated that evolutionary classification can significantly inform decision making in breeding and genetic manipulation in crop species as demonstrated in SAL mutant wheat field trials.

Abstract

High-throughput sequencing has generated over 246 million protein sequences, yet fewer than 0.3% have experimental validation. This gap is acute in plants, where frequent whole-genome duplication creates paralogues that sequence-based methods cannot reliably classify as functionally equivalent or divergent.

This thesis applies a progression of computational classification methods to plant enzymes. Structural modelling of phytoene synthase variants elucidated mechanisms of enzyme dysfunction. *Evolocity* network analysis of *DXO1* and *eIF4E* identified angiosperm-specific neo-functionalisation events. Sequence similarity networks combined with molecular phylogenetics classified the nucleotide phosphatase family into three distinct groups (SAL, AHL, CNP) and predicted functional divergence between wheat SAL paralogues, subsequently validated by field trials showing opposing yield effects.

Each method revealed limitations motivating the next, culminating in a novel approach comparing Fourier-decomposed active site electrostatic fields across 506 enzymes. Since catalysis depends on the electrostatic environment governing substrate binding, this captures functional information that sequence and structure cannot provide. Electrostatic analysis detected functional relationships invisible to other methods where Algal AHL and Ancient CNP clustered with SAL despite phylogenetic divergence, while AHL sub-clades showed regulatory divergence despite sequence conservation.

List of Publications

Publications arising from this thesis

Published

- Bowerman, A. F., Moore, M., Yadav, A., Zhang, J., Mortimer, M. D., Plšková, Z., Tee, E. E., Au, E. K., Collinge, D. P., Estavillo, G. M., Howitt, C. A., Chan, K. X., Rebetzke, G. J., & Pogson, B. J. (2025). Modulation of SAL retrograde signalling promotes yield and water productivity responses in dynamic field environments. *New Phytologist*. Advance online publication. <https://doi.org/10.1111/nph.7057>
- Hou, X., Alagoz, Y., Welsch, R., Mortimer, M. D., Pogson, B. J., & Cazzonelli, C. I. (2024). Reducing PHYTOENE SYNTHASE activity fine-tunes the abundance of a cis-carotene-derived signal that regulates the PIF3/HY5 module and plastid biogenesis. *Journal of Experimental Botany*, 75(4), 1187–1204. <https://doi.org/10.1093/jxb/erad443>

Under Review

- Ganguly, D. R., Yu, X., Mortimer, M. D., Moore, M., Khambalkar, P., Pogson, B. J., & Gregory, B. D. (2025). *Arabidopsis DXO1 is an evolutionarily diverged homolog that impacts ribosome loading and mRNA surveillance* (bioRxiv Preprint). <https://doi.org/10.1101/2025.10.29.684918>

Submitted to Proceedings of the National Academy of Sciences of the United States of America (PNAS)

In preparation

- Evolution of plant Eukaryotic Translation Initiation Factor Isoform 4E

- The molecular evolution of plant nucleotide phosphatases and functional prediction using an electrostatic similarity network

Thesis Roadmap

This thesis develops computational methods for classifying proteins when sequence and structural similarity fail to capture functional relationships. The thesis follows a methodological progression across three results chapters.

This thesis is written in the style of a thesis by publication. Chapter 3 comprises of work from one published paper, one paper under review, and one paper in preparation. Chapter 4 includes one published paper alongside unpublished foundational work. Chapter 5 is being prepared for publication and the paper will incorporate the unpublished foundational work from Chapter 4. My publishing contributions and roles are detailed at the start of each results chapter.

Each chapter builds on the preceding work: Chapter 5 uses the nucleotide phosphatase dataset generated in Chapter 4, and addresses limitations identified in both earlier chapters. Chapter 5 serves as both the culminating methods chapter and the thesis synthesis; its Discussion frames the chapter findings within the context of the overall thesis.

Chapter guide

Below is a detailed guide to the content of each chapter and considerations for reading:

Chapter 1: *Overview* provides a brief context for protein classification challenges and positions this thesis within the field, rather than a comprehensive literature review. Detailed introductions for specific topics are provided in the relevant results chapters.

Chapter 2: *Methods* describes established computational methods used across Chapters 3 and 4, including sequence database searching, alignment, similarity network construction, and phylogenetic inference. Chapter 5 contains its own Methods section specific to the novel electrostatic analysis pipeline.

Chapter 3: *Sequence and structural approaches to protein classification* presents three case studies: structural modelling of phytoene synthase mutations, and

evolutionary analysis of decapping exoribonuclease and eukaryotic translation initiation factors. These demonstrate current computational approaches while revealing characteristic limitations that motivate Chapter 5.

Chapter 4: *Sequence-based classification of plant nucleotide phosphatases* establishes the evolutionary framework for the plant nucleotide phosphatase family using sequence similarity networks and phylogenetic inference. It presents a crop field trial validating the classification approach and demonstrating the value of detailed classification for agricultural and biotechnology applications. The dataset analysed in Chapter 5 was generated in this chapter.

Chapter 5: *Enzyme active site electrostatic field comparison and similarity network* develops and validates a novel computational method for comparing enzyme active site electrostatics. The Results section presents both methodological development and biological findings, as the iterative refinement of the method is itself a primary contribution. The Discussion synthesises all thesis findings. Internal links in blue have been provided to take you to referenced sections or terms, when reading as a PDF document.

Appendix A: *Computational pipeline* provide the technical aspects of how the final method was implemented. This invoked extensive work including critical, but not directly biology-related, strategies to reduce run-times of several years to a few days.

Contents

Declaration of Originality	i
Acknowledgements	ii
Abstract	iv
List of Publications	vi
Roadmap	viii
List of Abbreviations	xvi
Definitions	xviii
List of Figures	xix
List of Tables	xxi
1 Thesis Overview	1
1.1 Introduction	1
1.2 Approaches to protein classification	2
1.2.1 Defining protein classification	2
1.2.2 Computational methods for protein classification	3
1.2.3 The impact of expanding data	7
1.3 Classification challenges in plant biology	8
1.3.1 Orthologues, paralogues, and gene duplication	8
1.3.2 Whole-genome duplication in plants	9
1.3.3 Limited experimental annotation in plants	10
1.4 Thesis scope and structure	10
1.4.1 Chapter summaries	11
1.5 Hypothesis and aims	12

2	Methods	13
2.1	Sequence dataset generation	13
2.1.1	Database searching	13
2.1.2	Redundancy reduction and quality filtering	14
2.1.3	Case-study specific parameters	14
2.2	Sequence similarity network analysis	16
2.2.1	Network construction	16
2.2.2	Network visualisation and clustering	16
2.2.3	Cluster identification and validation	16
2.2.4	Case-study specific parameters	16
2.3	Multiple sequence alignment	19
2.3.1	Alignment algorithms	19
2.3.2	Alignment optimisation	19
2.3.3	Profile alignments	19
2.3.4	Manual curation	20
2.3.5	Transit peptide removal	20
2.3.6	Case-study specific parameters	20
2.4	Phylogenetic inference	21
2.4.1	Evolutionary model selection	21
2.4.2	Maximum likelihood tree inference	22
2.4.3	Dataset reduction for outgroup clades	22
2.4.4	Statistical validation	23
2.4.5	Consensus tree construction and final tree selection	23
2.4.6	Tree rooting	23
2.4.7	Case-study specific parameters	23
2.5	Evolutionary velocity analysis	24
2.5.1	Network construction and velocity embedding	24
2.5.2	Case-study specific parameters	24
2.6	Structural modelling	25
2.6.1	AlphaFold2 and ColabFold	25
2.6.2	ESMFold	25
2.6.3	Model validation	25
2.6.4	Case-study specific parameters	25
2.7	Sequence logo analysis	26
2.8	Functional annotation	26
2.8.1	Subcellular localisation prediction	26
2.8.2	Co-expression analysis	26
2.8.3	Gene ontology enrichment	27
2.8.4	Tissue and condition-specific expression	27

2.9	Statistical methods	27
2.9.1	Comparison of sequence properties between clades . . .	27
2.9.2	Conservation of translation initiation factor isoforms . .	28
2.9.3	Serine-rich domain identification	28
2.10	Computational tools and software	29
2.10.1	Code and data availability	29
3	Sequence and structural approaches to protein classification	30
3.1	Introduction	32
3.1.1	Phytoene synthase	33
3.1.2	Decapping exoribonuclease	34
3.1.3	Eukaryotic translation initiation factors	35
3.2	Results	37
3.2.1	Pytoene synthase	37
3.2.2	Decapping exoribonuclease	40
3.2.3	Eukaryotic Translation Initiation Factor Isoform 4E . . .	44
3.3	Discussion	49
3.3.1	Technical findings	49
3.3.2	Biological findings	52
3.3.3	Experimental validation across case studies	60
3.3.4	Implications for protein classification methodology . . .	63
4	Sequence-based classification of plant nucleotide phosphatases	65
4.1	Publishing contribution	65
4.1.1	Significance statement	65
4.2	Introduction	66
4.3	Results	70
4.3.1	Classification of the plant nucleotide phosphatase family	70
4.3.2	Phylogenetic Classification of Wheat SAL Homologues .	75
4.3.3	Expression and Co-expression Analysis Support Functional Divergence	78
4.3.4	Field Performance of Targeted SAL Deletions	78
4.3.5	<i>Arabidopsis thaliana</i> nucleotide phosphatase expression and Gene Ontology	79
4.4	Discussion	82
4.4.1	Technical findings	82
4.4.2	Biological findings	83
4.4.3	<i>In vivo</i> validation of sequence-based classification	87
4.4.4	Gene dosage sensitivity in SAL evolution	88
4.4.5	Limitations of sequence-based approaches	95

4.4.6	Testable hypotheses	95
5	Enzyme active site electrostatic field comparison and similarity network	97
5.0.1	How this chapter is structured	97
5.0.2	Significance statement	97
5.1	Introduction	98
5.1.1	Why Electrostatics	99
5.1.2	Technical Advances Open Up Opportunity	101
5.1.3	Improving nucleotide phosphatase classification	110
5.1.4	Hypothesis	112
5.1.5	Data source	112
5.2	Methods	113
5.2.1	Computer system	113
5.2.2	Code availability	113
5.2.3	Structural modelling	113
5.2.4	Electrostatic calculations	114
5.2.5	Analysis methods	117
5.2.6	Statistical Significance Assessment	120
5.2.7	Similarity networks	120
5.3	Results	120
5.3.1	Approach 1: Comparisons of 2-dimensional image active site surface charge	120
5.3.2	Approach 2: Comparisons of three-dimensional active site electrostatic maps	127
5.3.3	Approach 3: Fourier decomposition of three-dimensional active site electrostatic maps	134
5.4	Discussion	168
5.4.1	Summary of key findings	168
5.4.2	Methodological progression from sequence to electrostatic classification	169
5.4.3	Biological interpretation of clustering patterns: Functional continuity of PAP catabolism across Viridiplantae	174
5.4.4	What frequency bands reveal about functional encoding	187
5.4.5	Relationship to existing methods	189
5.4.6	Limitations and opportunities for refinement	192
5.4.7	Generalisability beyond nucleotide phosphatases	200
5.4.8	Implications for enzyme classification	201
5.4.9	Future directions	201

5.4.10 Conclusion	204
A Computational Pipeline	206
B Python structural alignment and similarity network scripts	217
B.1 PyMol structural alignment	217
B.2 Active site coordinates	218
B.3 Structural similarity network construction	219
Bibliography	222

List of Abbreviations

A. thaliana *Arabidopsis thaliana* (L.) Heynh.

E. coli *Escherichia coli*.

S. cerevisiae *Saccharomyces cerevisiae*.

T. aestivum *Triticum aestivum* (wheat) (L.).

ccr2 carotenoid and chloroplast regulation 2.

aa amino acids.

AF2 AlphaFold2.

AHL ARABIDOPSIS HALOTOLERANCE 2-LIKE.

AMP Adenosine MonoPhosphate.

APBS Adaptive Poisson-Boltzmann Solver.

ARI Adjusted Rand Index.

BIC Bayesian Information Criterion.

Bit-score Bit-score.

BLAST Basic Local Alignment Search Tool.

CASP Critical Assessment of protein Structure Prediction.

CNP CHLOROPLASTIC NUCLEOTIDE PHOSPHATASE-LIKE.

CPU Central Processing Unit.

CRTISO CAROTENOID ISOMERASE.

DXO1 DECAPPING EXORIBONUCLEASE 1.

E-value Expect value.

eIF(ISO)4E EUKARYOTIC TRANSLATION INITIATION FACTOR ISOFORM 4E.

eIF4E EUKARYOTIC TRANSLATION INITIATION FACTOR 4E.

EMD Earth Mover's Distance.

Evolocity Evolutionary velocity with protein language models.

FDR False Discover Rate.

FFT Fast Fourier Transform.

GGPP geranylgeranyl diphosphate.

GGPPS11 GGPP SYNTHASE 11.

GPU Graphical Processing Unit.

HmmScan Hidden Markov Model Scan.

KNN K-Nearest Neighbours.

LGT Lateral Gene Transfer.

MAFFT Multiple Alignment using Fast Fourier Transform.

MCC Matthews Correlation Coefficient.

MIP Maximum Intensity Projection.

MSA Multiple Sequence Alignment.

NCBI National Center for Biotechnology Information.

OR ORANGE protein.

PAP 3'-phosphoadenosine 5'-phosphate.

PDB Protein Data Bank.

PDB2PQR PDB to PQR converter).

PERMANOVA Permutational Multivariate Analysis of Variance.

Pfam Pfam Protein Families.

PIPSA Protein Interaction Property Similarity Analysis.

pLDDT predicted local distance difference test.

PSY PHYTOENE SYNTHASE.

RAM random access memory.

RMSD Root-Mean-Square Deviation.

ScHAL2 *Saccharomyces cerevisiae* HALOTOLOERANCE-2.

SSN Sequence Similarity Network.

TaSAL1 *Triticum aestivum* L. SAL1.

TaSAL2 *Triticum aestivum* L. SAL2.

UMAP Uniform Manifold Approximate Projection.

WGD Whole Genome Duplication.

Definitions

function Protein function is used broadly in this work. It include molecular level function including, but not limited to catalytic, structural, regulatory. It also includes *in plantae* roles including physiological, developmental and metabolic systems roles..

Sinkhorn An iterative algorithm that efficiently computes the minimal distance between probability distributions.

wet-lab Non-computational experiments that include *in vitro*, *in vivo*, and *in plantae* experiments.

List of Figures

2.1	PSY sequence similarity network analysis	17
3.1	PSY mutation site conservation analysis	37
3.2	PSY structural model validation	38
3.3	Structural mapping of PSY mutations	39
3.4	Sequence information for Rai1/Dxo1/DXO homologues	41
3.5	Angiosperms harbour a divergent DXO1 protein characterised by a serine-rich N-terminal region	42
3.6	Network analysis of eIF4E and eIF(ISO)4E	45
4.1	Sequence similarity network (SSN) of SAL-homologue BLAST results.	72
4.2	400 phylogeny quality score distribution	73
4.3	Plant nucleotide phosphatase circular layout phylogeny.	74
4.4	AHL clades and representative structures	75
4.5	CNP putative domain swap	76
4.6	Wheat SAL gene diversity	77
4.7	AtAHL, AtSAL1, and AtCNP tissue expression	79
4.8	AtSAL1, AtAHL, AtCNP expression (mRNA) under drought.	80
4.9	Cladogram illustrating alternative topologies for SAL–AHL di- vergence	84
5.1	Representative electrostatic map images	100
5.2	Cropped AtSAL1 active site	115
5.3	AtSAL1 crystal structure–AlphaFold2 structural model compar- ison	122
5.4	Distribution of structural pLDDT scores	123
5.5	Example approach 1 active site images	123
5.6	Earth Mover’s Distance diagram	125
5.7	Active site image similarity network	127
5.8	AtSAL1-homologue pairwise comparison	132

5.9	Active site electrostatic distance heatmap and dendrogram using the Sinkhorn algorithm	133
5.10	Sinkhorn distance distribution	133
5.11	Approach 2 Sinkhorn derived similarity network	134
5.12	Electrostatic Fourier decomposition conceptual visualisation . .	135
5.13	Approach 3 computational process	139
5.14	Grid size parameter sweep	146
5.15	15 Band Spearman correlation matrix	149
5.16	Fourier decomposition dendrogram heatmap of frequency band 4	152
5.17	Fourier decomposed active site electrostatic similarity network, band 4, paralogue clustering	153
5.18	Similarity network comparison	155
5.19	Similarity network comparison with algal sequences	156
5.20	Clustering robustness under progressive noise	157
5.21	Paralogue distance comparison control	158
5.22	Fourier decomposed active site electrostatic similarity network, band 5, taxa clustering	160
5.23	Fourier decomposed active site electrostatic similarity network, band 4, localisation clustering	164
5.24	Functional evolution of PAP catabolism in plant nucleotide phosphatases	179

List of Tables

2.1	<i>Arabidopsis thaliana</i> (L.) Heynh (<i>A. thaliana</i>) nucleotide phosphatase homologue reference lengths	15
2.2	Computational tools used for sequence-based classification	29
3.1	Number of DXO1 homologues retrieved per Kingdom	43
3.2	Non-plant DXO1 homologues with serine-rich N-terminals. . . .	43
3.3	Distribution of EUKARYOTIC TRANSLATION INITIATION FACTOR 4E (eIF4E) and EUKARYOTIC TRANSLATION INITIATION FACTOR ISOFORM 4E (eIF(ISO)4E) across plant orders	47
3.4	Statistical analysis of eIF(ISO)4E retention across angiosperm or- ders	48
4.1	Welch <i>t</i> -test using Cohen's <i>d</i> comparing the outparalogous AR- ABIDOPSIS HALOTOLERANCE 2-LIKE (AHL) clades, <i>n</i> = 84*	74
4.2	Welch <i>t</i> -test using Cohen's <i>d</i> comparing the outparalogous CNP clades, <i>n</i> = 73*	76
4.3	Yield, biomass, and water productivity.	78
4.4	Gene ontology analysis of genes co-expressed with AtSAL1, AtAHL, and AtCNP	81
5.1	Approach 2 paralogue agreement results	131
5.2	PERMANOVA R^2 Band 4 and pH condition for paralogue classi- fication at different ϵ multiplier values.	147
5.3	PERMANOVA R^2 values by band and pH condition for para- logue classification.	153
5.4	PERMANOVA R^2 values by band and pH condition for taxa clas- sification.	159
5.5	AHL PERMANOVA R^2 values by band and pH condition for taxa classification (<i>n</i> = 187).	161
5.6	CNP PERMANOVA R^2 values by band and pH condition for taxa classification (<i>n</i> = 159).	161

5.7	SAL PERMANOVA R^2 values by band and pH condition for taxa classification (n = 160).	161
5.8	PERMANOVA R^2 values by band and pH condition for localisation classification (n = 506).	165
5.9	Proportion of plastid localisation by paralogue (n = 506).	165
5.10	PERMANOVA R^2 values by band and pH condition for combined paralogue, taxa and localisation classification (n = 506).	166
5.11	PERMANOVA R^2 values by band and paralogue for pH classification (n = 506).	167
5.12	PERMANOVA R^2 values by band and AHL sub-clade for pH classification (n = 506).	168

Chapter 1

Thesis Overview

This chapter provides a brief overview of approaches to protein classification and the specific challenges that arise in plant biology. Each subsequent chapter contains a detailed introduction to its specific methods and biological context.

1.1 Introduction

The UniProt database now contains over 246 million protein sequences, increasing at a rate that far outstrips our capacity for experimental characterisation; fewer than 0.3% sequences have experimentally validated functional annotations

(Blum et al., 2025). Computational approaches to protein classification are therefore essential for interpreting data at this scale. This thesis explores several sequence-based approaches for protein classification and putative annotation, then builds on these methods to develop a novel computational approach by comparing enzyme electrostatic fields, using plant nucleotide phosphatases as a validation system throughout.

Without reliable functional classification, the value of the considerable and ever increasing volume of sequences data is significantly reduced. Understanding how organisms respond to environmental stress, how metabolic pathways are regulated, or the function of uncharacterised proteins depends on accurate functional annotation. This thesis focuses on exemplar plant enzymes from a basic science and agronomic perspective.

1.2 Approaches to protein classification

1.2.1 Defining protein classification

Protein classification is the organisation of proteins into groups based on shared properties. These properties may include sequence or structural similarity, physiological role, evolutionary relationships, or biochemical function. Classification enables correlation of functional annotations from characterised to uncharacterised proteins, reveals evolutionary relationships between protein families, and provides frameworks for understanding how sequence variation relates to functional diversity. Importantly, it provides context for understanding metabolic systems and the molecular basis for physiological insights where experimental characterisation of proteins is needed to understand the system but experiments are not possible or are out of reach.

Classification is related to, but distinct from, annotation and function prediction. Annotation assigns specific labels to proteins, such as Gene Ontology terms or Enzyme Commission (EC) numbers. Functional prediction infers what a protein does based on computational analysis. Classification provides the groupings within which annotation transfer and function prediction operate. The reliability of annotation and prediction therefore depends on the accuracy of the underlying classification.

The sequence-structure-function paradigm provides the conceptual foundation for most classification approaches ([Whisstock and Lesk, 2003](#); [Lin et al., 2024](#)). This paradigm asserts that amino acid sequence determines three-dimensional structure through the protein folding process, and that structure in turn enables biological function by positioning key residues and creating binding sites. Homology-based classification relies on this paradigm where proteins that share sufficient sequence similarity are inferred to share structure and therefore function.

Enzymes with identical folds can catalyse unrelated reactions, exemplified in the foundational work on the TIM barrel fold ([Wierenga \(2001\)](#)), reviewed in [Romero-Romero et al. \(2021\)](#), while other proteins with different structures can converge on the same function ([Omelchenko et al., 2010](#)), suggesting this paradigm is incomplete. The relationship between sequence, structure, and function is correlative rather than deterministic, and the strength of this correlation varies across protein families and evolutionary distances.

1.2.2 Computational methods for protein classification

Experimental characterisation of protein function remains the gold standard for functional annotation, but cannot keep pace with the rate of sequence discovery (The UniProt Consortium, 2015). Computational methods enable classification and annotation transfer from characterised proteins to the vast majority that lack experimental data. Without computational classification, the functional context of most genomic data would remain unknown (Radivojac et al., 2013; Lin et al., 2024).

Computational annotation carries inherent uncertainty, and errors can propagate throughout databases. Analysis of enzyme superfamilies has revealed misannotation rates ranging from 5% to 63% depending on the protein family, with some families exceeding 80% error rates (Schnoes et al., 2009). These errors are predominantly over-predictions, where proteins are assigned functions more specific than the evidence supports. A single misannotated founding member of a protein family can propagate errors to dozens of subsequently annotated sequences, each then propagating the error further (Schnoes et al., 2009). Understanding the strengths and limitations of different classification methods is therefore important for assessing annotation reliability, and is critical for maintaining database integrity.

Sequence similarity methods

Sequence homology methods have been instrumental in protein classification. Methods such as BLAST and PSI-BLAST, introduced by Altschul et al. (1997), detect homology through local sequence alignment and remain widely used for identifying related proteins (Lin et al., 2024). Profile-based methods have improved classification by identifying protein family-specific sequence patterns. Pfam hidden Markov models now classify approximately 77% of UniProtKB sequences into protein families (Mistry et al., 2021), a crucial service but limited in many contexts as protein families themselves can be functionally very broad.

Sequence-based classification is most reliable when query and reference sequences share high identity and the protein family is under strong selective pressure. Above approximately 60% sequence identity, annotation accuracy for general enzyme classes approaches 80% (Tian and Skolnick, 2003; Yu et al., 2023). Reliability decreases for fine-grained functional predictions and distant homologues. Below 40% sequence identity, predictions of substrate specificity become unreliable (Tian and Skolnick, 2003; Yu et al., 2023). The relationship between sequence similarity and functional conservation is not linear, and the

threshold at which functional inference becomes unreliable varies between protein families.

Phylogenetic methods

Molecular phylogenetics extends protein classification by applying evolutionary models to infer proteins' evolutionary relationships. Phylogenetic reconstruction using maximum likelihood or Bayesian approaches provides evolutionary context that sequence similarity alone cannot capture ([Nascimento et al., 2017](#)). Critically, phylogenetic methods can distinguish orthologues from paralogues and identify gene duplication events. Identifying orthologues versus paralogues is essential for functional classification, as the former is defined by having the same function between different species, while the latter requires a duplication event with no functional equivalence being able to be proscribed. Importantly sequence similarity scores alone cannot reliably determine orthology.

Phylogenetic methods can trace the evolutionary fate of duplicated gene copies, showing where duplications likely have occurred and where proteins are likely synonymous between species. However, phylogenetic reconstruction becomes computationally demanding for large datasets, and the accuracy of inferred relationships depends on adequate taxon sampling and appropriate evolutionary models ([Nabhan and Sarkar, 2012](#); [Young and Gillung, 2020](#)).

Sequence similarity networks

Sequence similarity networks provide an alternative approach for visualising relationships within large protein families ([Atkinson et al., 2009](#)). These networks represent proteins as nodes and sequence similarities as edges, with edge weights reflecting alignment scores. Clustering algorithms applied to these networks can identify groups of related proteins without requiring explicit phylogenetic reconstruction.

Sequence similarity networks are particularly useful for exploring large superfamilies where phylogenetic trees become unwieldy. ([Atkinson et al., 2009](#); [Zallot et al., 2019](#)) They enable rapid identification of distinct protein groups based on sequence similarity, and can reveal relationships that are difficult to discern from trees alone. These networks benefit from being computationally cheap, can handle large datasets, and allows the data to be explored through interaction with the networks – such as colouring nodes by known features including taxonomy, biochemical function, or can be integrated with other prediction datasets ([Atkinson et al., 2009](#); [Zallot et al., 2019](#)). However, similar-

ity networks complement rather than replace phylogenetic analysis. Networks show similarity relationships but do not distinguish orthology from paralogy, and the clusters they identify require validation through other methods.

Structural methods

Structure is more conserved than sequence during evolution, retaining recognisable similarity even when sequences have diverged beyond recognition (foundational work in [Chothia and Lesk \(1986\)](#) and developed in [Illergård et al. \(2009\)](#), continued in [Konaté et al. \(2019\)](#)). Structural comparison can therefore detect relationships that sequence methods miss. Classification systems such as SCOP and CATH organise proteins by structural similarity, grouping proteins into folds, superfamilies, and families based on three-dimensional architecture ([Andreeva et al., 2020](#)).

Until recently, accurate structure determination required months to years per protein through X-ray crystallography, nuclear magnetic resonance spectroscopy or Cryogenic electron microscopy. This constraint limited structural classification to the small fraction of proteins with experimentally determined structures. The machine learning structural modelling tool [AlphaFold2 \(AF2\)](#) changed this by predicting protein structures with accuracy approaching experimental resolution ([Jumper et al., 2021](#)). The AlphaFold Protein Structure Database now contains predictions for over 200 million proteins ([Varadi et al., 2024](#)), enabling structure-based classification at massive scales.

However, structural similarity does not guarantee functional similarity which is exemplified by the TIM barrel fold. The TIM barrel fold is one of the most common enzyme architectures, it occurs in dozens of enzyme families catalysing unrelated reactions ([Nagano et al., 2002](#); [Romero-Romero et al., 2021](#)). Conversely, different structures can converge on the same function ([Ribeiro et al., 2023](#)), as demonstrated by the serine proteases trypsin and subtilisin, which have different structural folds yet catalyse the same reaction ([Dodson and Wlodawer, 1998](#)). This is a core challenge of classification through structural similarity as structure prediction tools such as [AF2](#) predict what a protein looks like but provide no direct information about substrate preference, catalytic efficiency, or physiological context.

Neither sequence nor structural methods directly capture electrostatic properties which govern enzyme catalysis ([Warshel, 1998](#); [Warshel et al., 2006](#); [Ruiz-Pernía et al., 2024](#)). This is a key limitation of the sequence-structure-function paradigm, where it is electrostatics, not structure, from which catalytic function is directly derived from. Enzymes accelerate chemical reactions by factors

of 10^6 to 10^{17} compared to uncatalysed rates (Wolfenden and Snider, 2001; Richard, 2019). The conventional explanation for this catalytic power focuses on geometric complementarity, where enzymes position substrates and catalytic residues in precise spatial arrangements that favour the reaction. While geometry is essential, enzymes with nearly identical active site geometries can differ markedly in catalytic efficiency and substrate preference (Warshel et al., 2006; Fried and Boxer, 2017). The work of Arie Warshel established that electrostatic stabilisation of the transition state is the dominant contribution to enzymatic rate enhancement (Warshel, 1998; Warshel et al., 2006). Enzymes create pre-organised electric fields that preferentially stabilise the charge distribution of the transition state relative to reactants. Experimental measurements of electric fields within enzyme active sites have confirmed Warshel's assertion that local field strength correlates with catalytic efficiency (Fried and Boxer, 2017).

Beyond the active site, electrostatic fields guide substrates toward binding sites, accelerating substrate binding beyond diffusion-limited rates (Wade et al., 1998; Zhou and Pang, 2018). Electrostatic steering contributes to both catalytic efficiency and substrate specificity, as proteins processing charged substrates typically display complementary surface charge distributions that favour productive encounters. Furthermore, electrostatic properties respond to environmental conditions that sequence and structure cannot capture. Plant cells contain compartments spanning pH values from approximately 4.5 in the apoplast to above 8 in the peroxisome (Garcia-Moreno, 2009), and enzymes localised to different compartments often display pH optima matching their environment. The chloroplast stroma drops to approximately pH 7.0 when photosynthesis ceases but rises above 8.0 during active photosynthesis (Heldt et al., 1973; Trinh and Masuda, 2022), altering the electrostatic state of stromal enzymes without any change to their sequence or structure.

The plant enzyme ribulose-1,5-bisphosphate carboxylase/oxygenase (RuBisCO) illustrates how electrostatic properties can distinguish functionally divergent enzymes that sequence and structure methods classify together. C3 and C4 plant variants share highly conserved active site structures, yet display different kinetic properties adapted to their distinct carbon fixation contexts (foundational work with Spreitzer and Salvucci (2002) and reviewed by Sharwood (2017)). Recent structural analysis demonstrated that positively charged cavities surrounding the active site correlate with substrate specificity across divergent RuBisCO forms, with cavity surface area explaining approximately 60% of specificity variance (Poudel et al., 2020). If electrostatic similarity encodes functional information not captured by sequence or structure alone, direct comparison of these properties could complement existing classification ap-

proaches. Chapter 5 examines the methodological requirements for such comparisons.

Emerging computational methods

Recent advances in machine learning have introduced new approaches to protein classification. Methods such as DeepFRI predict Gene Ontology terms directly from protein structures using graph neural networks trained on experimentally annotated proteins (Glgorijević et al., 2021). These approaches can make predictions for proteins lacking close homologues in training data, potentially extending classification beyond what sequence similarity methods can achieve.

Protein language models represent another emerging approach. Models such as ESM learn representations of protein sequences by training on large sequence databases, capturing patterns of evolutionary constraint without explicit supervision (Lin et al., 2023). These learned representations can be used for various downstream tasks including function prediction and classification.

Evolutionary velocity with protein language models (Evolocity) uses protein language models to analyse evolutionary trajectories within protein families (Hie et al., 2022). By constructing vector fields of mutational likelihood across sequence space, Evolocity can predict the directionality and rate of evolutionary change and identify distinct evolutionary lineages. Its pseudo-time analysis tool 'backtracks' along this evolutionary trajectory in order to identify the putative root of the network, in a similar manner as a phylogenetic tree. Evolocity scales to large sequence datasets where traditional phylogenetic reconstruction becomes computationally impractical. These emerging methods show promise but remain under active development, and their performance relative to established approaches requires continued evaluation. In particular, most machine learning approaches require significant amounts of labelled data to predict what those labels are in other proteins. This presents a challenge as only a small percentage of the available training data has experimental validation, and as the training data comes from existing biological datasets is victim to the sampling bias and propagated errors of those datasets.

1.2.3 The impact of expanding data

The scale of available sequence data has grown exponentially (The UniProt Consortium, 2015). Metagenomic sequencing continues to reveal protein diversity from uncultured organisms, while genome sequencing of non-model species expands coverage across the tree of life.

The availability of predicted structures through AF2 represents a parallel expansion. Structure-based classification, previously limited to the few percent of proteins with experimental structures, can now be applied to essentially any protein with a known sequence. This creates opportunities for detecting distant relationships and validating sequence-based classifications.

However, more data does not automatically improve classification. The challenge still remains to ascertain which proteins share function, and is limited by how reliably computational methods can identify functional similarity. Expanded data increases the power of comparative approaches but also increases the diversity of proteins requiring classification, including many from poorly characterised organisms and protein families. Mislabelled data propagates errors (see 1.2.2) which contaminate the output of many of these methods.

These general challenges are particularly acute in plant biology, where genome complexity and limited experimental data create specific difficulties for protein classification. This is introduced below.

1.3 Classification challenges in plant biology

1.3.1 Orthologues, paralogues, and gene duplication

Protein classification in essence is about labelling uncharacterised proteins based on correlation or similarity with other proteins and depends on understanding the evolutionary relationships between genes (Gabaldón and Koonin, 2013; Altenhoff et al., 2019). Two genes in different species that descended from a common ancestral gene through speciation are orthologues, while those from within the same genome that originate from gene duplication events are paralogues. As briefly discussed in section 1.2.2 in the phylogenetics paragraph, orthologues typically retain similar function because they have been subject to similar selective pressures since diverging. Two genes within the same species, or in different species, that arose through gene duplication are paralogues. Paralogues may retain similar function or may diverge, depending on the selective pressures acting on each copy following duplication.

Gene duplication occurs through several mechanisms. Whole-genome duplication creates copies of all genes simultaneously, typically through errors in cell division or hybridisation between species (Van de Peer et al., 2017; Qiao et al., 2019). Segmental duplication copies chromosomal regions containing one or more genes. Tandem duplication creates adjacent copies of individual genes through unequal crossing over during meiosis. Each mechanism produces paralogues with different genomic contexts and potentially different evol-

utionary trajectories.

Following duplication, the fate of gene copies varies as both copies may retain the original function, providing redundancy. One copy may acquire a new function – neo-functionalisation – while the other retains the ancestral role. The ancestral function may be partitioned between copies – a process called sub-functionalisation – with each copy becoming specialised for a subset of the original function. Alternatively, one copy may accumulate deleterious mutations and become a non-functional pseudogene. The balance between these outcomes depends on population size, the strength of selection, and the functional constraints on the protein (Birchler and Yang, 2022).

The distinction between orthologues and paralogues has direct consequences for functional classification. Annotation transfer between orthologues is generally reliable because orthologues are expected to share function. Annotation transfer between paralogues is uncertain because paralogues may have diverged functionally even while retaining high sequence similarity. Distinguishing these relationships is therefore essential for accurate classification, yet sequence similarity alone cannot distinguish orthology from paralogy.

1.3.2 Whole-genome duplication in plants

Plant genomes have undergone recurrent whole-genome duplication events throughout their evolutionary history. Plants tolerate genome multiplication better than most animals and fungi (Albertin and Marullo, 2012; Van de Peer et al., 2017). Approximately 70% of angiosperms have undergone at least one whole-genome duplication event since their emergence (Masterson, 1994), and roughly 15% of speciation events involve increases in ploidy (Van de Peer et al., 2017).

These ancient whole-genome duplication events have left lasting signatures in plant genomes. For example, the gamma-polyploidy event, which occurred approximately 130–160 million years ago at the base of core eudicots, affected all descendants including major crop families (Jiao et al., 2012). More recent polyploidy events are superimposed on the more ancient duplication events, creating nested layers of paralogues with different ages and degrees of divergence. An estimated 64.5% of plant genes exist as plant-specific paralogues (Panchy et al., 2016).

This prevalence of gene duplication creates uncertainty in functional classification. Paralogues arising from whole-genome duplication initially have identical sequences and presumably identical function. Over evolutionary time, copies may diverge through neo-functionalisation or sub-functionalisation, or

one copy may be lost. The high sequence similarity between recently duplicated paralogues means that sequence-based classification methods group them together, even when their functions have begun to diverge. Conversely, ancient paralogues may have diverged sufficiently that their shared ancestry is no longer apparent from sequence comparison. This presents significant challenges for sequence-based approaches.

1.3.3 Limited experimental annotation in plants

The challenge of protein classification in plant biology is compounded by limited experimental data. Even the model organism *A. thaliana* has only approximately 7% of proteins with experimentally determined molecular functions (Reiser et al., 2024). This figure is notable given that Arabidopsis accounts for approximately 70,000 experimental Gene Ontology (GO) annotations, representing roughly 10% of all experimental GO annotations globally (de Crécy-lagard et al., 2022).

Experimental knowledge is heavily biased toward a small number of model organisms. Approximately 85% of global experimental annotations derive from just ten organisms, of which only one is a plant (de Crécy-lagard et al., 2022). For crop species and non-model plants, experimental characterisation is minimal, and annotation relies almost entirely on computational classification based on similarity to these few well-studied organisms.

This dependence on computational classification means that the limitations of classification methods have direct consequences for plant biology. If classification methods cannot distinguish functionally divergent paralogues, this uncertainty then propagates to every plant genome annotated using those methods. The prevalence of whole-genome duplication in plants means that paralogue classification is not an edge case but a core challenge affecting a substantial fraction of plant genes, and is the central challenge this thesis explores.

1.4 Thesis scope and structure

This thesis implements several sequence based approaches before developing a novel computational methodology for protein classification in plants, with particular focus on distinguishing functionally divergent paralogues. This integrates phylogenetic and network approaches, and methods for comparing enzyme electrostatic properties using Fourier decomposition. The work uses plant nucleotide phosphatases as a validation system through chapters four and five. This enzyme family includes the well-characterised AtSAL1 from

A. thaliana, which was identified through sequence homology with *Saccharomyces cerevisiae* HALOTOLOERANCE-2 (ScHAL2), a yeast salt tolerance gene (Murguía et al., 1995). AtSAL1 regulates chloroplast-to-nucleus retrograde signalling through its phosphatase activity (Estavillo et al., 2011; Chan et al., 2016), and it's family includes paralogues and related enzymes whose functional relationships remain unclear, providing classification problems that this thesis addresses.

1.4.1 Chapter summaries

First I describe the computational and experimental methods used throughout this thesis. This includes sequence retrieval and alignment, structural modelling with AF2, phylogenetic reconstruction, and statistical approaches for comparing protein properties across large datasets in chapter 2.

Next I apply established computational approaches to protein classification through three case studies in plant biology. The first examines how structural modelling can elucidate the effects of point mutations on enzyme function, using phytoene synthase as a model system. The second and third case studies apply Evolocify network analysis to classify protein variants and test evolutionary hypotheses for decapping exoribonuclease and translation initiation factors. This chapter discusses the application of structural modelling and sequence-based evolutionary analysis, including the technical capabilities and limitations that emerge from each case study. From this basis chapter 4 will explore the classification of nucleotide phosphatases using a Sequence Similarity Network (SSN) and phylogenetic analysis.

I then develop a sequence-based classification framework for plant nucleotide phosphatases. Chapter 4 integrates sequence similarity networks with phylogenetic inference to resolve the family into distinct groups and examines patterns of gene duplication and retention across plant lineages. The classification framework is tested against independent phenotypic data from wheat field trials, examining whether computational classification predicts functional outcomes relevant to crop improvement. The chapter addresses sequence similarity networks, phylogenetic reconstruction, and computational localisation prediction, including how these methods inform functional classification and where their limitations emerge. An approach that seeks to address these core limitations will be developed in chapter 5.

Finally I develop electrostatic field analysis as a complementary approach to protein classification, building on the insights gained from the limitations of the methods in the preceding chapters, using electrostatic comparison to

provide information beyond what sequence and structural methods capture. This chapter develops a Fourier decomposition approach for representing and comparing active site electrostatic fields and tests whether electrostatic similarity captures functional relationships not detected by sequence or structural comparison. It takes a novel approach to comparing the Fourier decomposed electrostatic maps through the use of optimal transport algorithms. The chapter covers electrostatic calculation methods, the technical challenges of comparing continuous scalar fields, and the requirements for this approach to complement existing classification methods.

The chapter ends with a Discussion integrating the findings and evaluates their comparative utility for protein classification in plants. This generates multiple detailed testable hypothesis for future, targeted, 'wet-lab' *in vivo* and *in vitro* work.

1.5 Hypothesis and aims

This thesis tests whether electrostatic similarity encodes functional relationships not directly assignable from direct sequence similarity. Comparing these properties directly should complement existing classification approaches for distinguishing functionally distinct paralogues.

This hypothesis tests whether these approaches are complementary to existing methods or can provide superior approaches for characterising functionally distinct homologues, providing a basis for inferring functional divergence that could be validated against kinetic data in future work.

The specific aims of this thesis are:

1. Examine whether sequence-based and structure-based approaches have characteristic limitations for functional classification of plant protein families.
2. Develop a sequence-based classification framework for plant nucleotide phosphatases that assigns uncharacterised homologues to functional groups.
3. Test whether computational classification can predict phenotypic outcomes using independent field trial data from wheat varieties.
4. Develop a computational methodology for comparing enzyme electrostatic properties at scale across homologous proteins.
5. Test whether electrostatic similarity captures functional relationships not detected by sequence or structural comparison.

Chapter 2

Methods

This chapter describes the computational methods used for sequence-based protein classification in Chapters 3 and 4. Methods are organised by methodology type, with case-study specific parameters detailed where applicable. Electrostatic field analysis methods are described separately in Chapter 5.

2.1 Sequence dataset generation

2.1.1 Database searching

Homologous sequences were identified through iterative database searching using protein Basic Local Alignment Search Tool (BLAST) (Altschul et al., 1990). Query sequences comprised characterised *A. thaliana* proteins relevant to each analysis, while searches were conducted against multiple databases depending on the taxonomic scope required:

- National Center for Biotechnology Information (NCBI) non-redundant protein database for broad taxonomic sampling
- Phytozome 12.1 (<https://phytozome.jgi.doe.gov/>) (Goodstein et al., 2012) for curated plant genome sequences
- China National GeneBank (CNCB) database of the One Thousand Plant Transcriptomes Initiative (<https://db.cngb.org/onekp/>) (Leebens-Mack et al., 2019) for broad sampling across Viridiplantae
- UniProt (The UniProt Consortium, 2023) for annotated protein sequences with taxonomic metadata

For targeted searches, TBLASTN (protein query against translated nucleotide database) was used with default parameters. Taxonomic information was retrieved from NCBI Entrez using the BioPython package (Cock et al., 2009).

2.1.2 Redundancy reduction and quality filtering

Initial datasets were filtered to remove redundancy and low-quality sequences. CD-HIT (Li and Godzik, 2006; Huang et al., 2010) was used to cluster sequences at specified identity thresholds, retaining a single representative from each cluster. The threshold selection balanced redundancy reduction against preservation of functionally relevant diversity.

Sequences were filtered by length to remove likely gene fragments and any mis-annotations. Length thresholds were set relative to characterised *A. thaliana* homologues, typically $\pm 30\%$ of the mature protein length (excluding transit peptides). This range accommodated natural length variation across Viridiplantae while excluding sequences too short to represent functional proteins or too long to be orthologous.

Duplicate sequences and those containing undefined residues (X characters) were removed programmatically using custom BioPython scripts (Cock et al., 2009).

2.1.3 Case-study specific parameters

Phytoene synthase

Putative PHYTOENE SYNTHASE (PSY) homologues were identified using protein-protein BLAST through Phytozome-13 (Goodstein et al., 2012) from the *A. thaliana* PSY sequence with the signal peptide removed. The search was against all Phytozome-13 non-redundant genomes with an Expect value (E-value) threshold of 10^{-1} , using the BLOSUM62 scoring matrix, default word length, and gaps allowed.

Decapping exoribonuclease

The amino acid sequence representing the primary isoform of *A. thaliana* DE-CAPPING EXORIBONUCLEASE 1 (DXO1) (AT4G17620.1) was used to identify its cognate Pfam Protein Families (Pfam) clan (CL0236) (Potter et al., 2018). Sequences within Pfam family PF08652 were used for analysis. Sequences underwent manual curation to incorporate taxonomic information (BioPython Entrez) and remove duplicates or those with ambiguous amino acids. Sequences with length greater than 2 standard deviations from the mean were removed (Figure 3.4A), as were sequences from Kingdoms with $n < 30$ (Table 3.1).

Eukaryotic translation initiation factor

Hidden Markov Model Scan (Hmmscan) (Potter et al., 2018) was used to identify the Pfam protein family from *A. thaliana* sequence At5g35620. The python BioServices package (Cokelaer et al., 2013) was used to download all sequences in Pfam family PF01652 from UniProt. Sequences with length less than 120 residues were filtered out, and duplicate sequences and those with ambiguous residues were removed.

Plant nucleotide phosphatases

Six *A. thaliana* nucleotide phosphatase sequences were used as seeds, see Table 2.1.

Each sequence was used in separate TBLASTN searches against Phytozome 12.1 and the CNGB One Thousand Plant Transcriptomes database (Leebens-Mack et al., 2019). This search strategy yielded 30,787 sequences. Length thresholds were set at $\pm 30\%$ of the mature protein length for each *A. thaliana* homologue (Table 2.1).

Table 2.1: *A. thaliana* nucleotide phosphatase homologue reference lengths

Homologue	TAIR accession	Length (amino acids)
AtAHL	At5G54390	373
AtSAL1	At5G63980	353
AtSAL2	At5G64000	347
AtSAL3	At5G63990	357
AtSAL4	At5G09290	345
AtCNP	At4G05090	397

Wheat SAL

Homologous sequences were retrieved using the UniProt web-based BLAST search against the AtSAL1 (NP 201203.3) sequence with parameters: blastp, E-value threshold 0.1, restricted to *Liliopsida*. Full sequences were retrieved from UniProt (The UniProt Consortium, 2023) using REST API access. AtSAL1 and other paralogous *A. thaliana* reference sequences were added to the dataset. Sequences were filtered to amino acid lengths between 250 and 600 residues. Existing wheat sequences were replaced with homologues identified from Gramine (Tello-Ruiz et al., 2022).

2.2 Sequence similarity network analysis

SSNs cluster proteins based on pairwise sequence similarity, enabling identification of orthologous relationships and discovery of previously unrecognised protein groups (Atkinson et al., 2009; Song et al., 2008). Networks were constructed and visualised using consistent methodology across all analyses.

2.2.1 Network construction

SSNs were generated using two approaches depending on dataset size and requirements:

EFI-EST pipeline For the decapping exoribonuclease and plant nucleotide phosphatase analyses, filtered sequence lists were submitted to the Enzyme Function Initiative Enzyme Similarity Tool (EFI-EST) (Gerlt et al., 2015) using the FASTA input option (Option C) with default parameters.

Local BLAST networks For datasets requiring greater control over parameters, all-versus-all BLAST networks were constructed using BLAST+ (NCBI-blast-2.12.0+) (Camacho et al., 2009). E-value thresholds were set according to dataset requirements, typically 10^{-10} .

2.2.2 Network visualisation and clustering

Networks were visualised in Cytoscape 3.7.2–3.9.1 (Shannon et al., 2003). Edge filtering was applied to retain relationships above specified similarity thresholds, with thresholds optimised for each dataset to achieve clear cluster separation while maintaining biologically meaningful connections. Clusters were generated using the Prefuse Force Directed OpenCL layout algorithm.

2.2.3 Cluster identification and validation

Clusters were identified using characterised *A. thaliana* sequences as controls. Sequence lists for each cluster were extracted using custom Python scripts, and FASTA files were generated with sequences annotated by cluster membership.

2.2.4 Case-study specific parameters

Phytoene synthase

The SSN was built using BLAST+ with an all-versus-all search and E-value threshold of 10^{-10} . The network (Figure 2.1) was visualised in Cytoscape 3.9.1 (Shannon et al., 2003) and clustered using the Prefuse Force Directed OpenCL layout. Edges with Bit-score (Bit-score) less than 595 were removed, and sequences from the cluster containing the *A. thaliana* PSY sequences were retained for alignment.

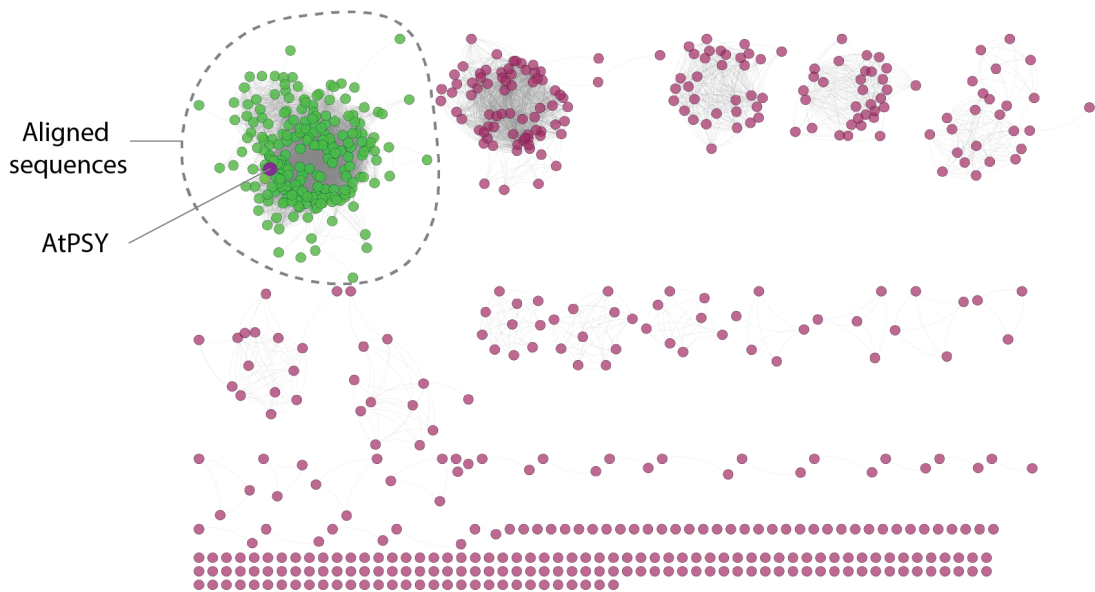


Figure 2.1: PSY sequence similarity network analysis.

Sequence similarity network of PSY homologues constructed using BLAST+ with E-value threshold of 10^{-10} . Network shows clustering of plant PSY sequences separate from other organisms, indicating plant-specific protein sequences. The plant clade was validated using the AtPSY sequence as a reference (purple). Putative plant PSY sequences are shown in green.

Decapping exoribonuclease

Sequences within Pfam clan CL0236 were used to generate a SSN using EFI-EST with $\geq 40\%$ sequence similarity threshold (Oberg et al., 2023). The node cluster containing AT4G17620.1 was used to attempt identification of a prokaryotic common ancestor; none was found. Analysis proceeded with sequences from Pfam family PF08652.

Eukaryotic translation initiation factor

An all-versus-all BLAST network was built using BLAST+ (NCBI-blast-2.12.0+) (Camacho et al., 2009) with an E-value threshold of 10^{-10} . The network was

visualised as a SSN in Cytoscape 3.9.1 (Shannon et al., 2003). *Bamfordvirae* (viruses), viral metagenome, marine metagenome, and duplicate ‘wolf’ sequences were removed. The SSN was filtered by percentage identity edges, removing those with values less than 52.5%, and the core cluster identified by *A. thaliana* reference sequences. This threshold maintained contiguous edges between all four domains. Clusters not connected to the core network were removed.

A plant-only sub-network was extracted and edges with percentage identity less than 54% were removed. The eIF(ISO)4E cluster was extracted into a sub-network and the node list exported to CSV to define eIF(ISO)4E sequences; the same procedure was applied for eIF4E. These labels were used for subsequent analysis.

Plant nucleotide phosphatases

The initial SSN was generated using EFI-EST with default parameters. Network edges were filtered to retain alignment scores greater than 70. A refined SSN was generated from the curated dataset, restricted to sequences with similarity scores greater than 10. In Cytoscape, edges with alignment scores below 35 were hidden. The final dataset was reduced to 500 (excluding outgroup) representative sequences using CD-HIT, comprising 160 SAL sequences, 185 AHL sequences, 155 CNP sequences, and 6 outgroup sequences.

Wheat SAL

An all-versus-all BLAST network was built using BLAST+ (NCBI-blast-2.12.0+) (Camacho et al., 2009) with an E-value threshold of 10^{-10} . The network was visualised as a SSN in Cytoscape 3.9.1 (Shannon et al., 2003). The SSN was filtered by percentage identity, removing edges with values less than 40%, until the core cluster identified by *A. thaliana* SAL1/2/3/4 reference sequences was separated from clusters containing paralogous *A. thaliana* reference sequences. Clusters not connected to the core network were removed. The *A. thaliana* reference sequences were then removed from the dataset. A randomised set of 11 sequences from the AHL cluster were added as an outgroup for phylogenetic reconstruction.

2.3 Multiple sequence alignment

2.3.1 Alignment algorithms

Two primary alignment algorithms were used:

T-COFFEE T-COFFEE (Notredame et al., 2000) was used with regression alignment (Garriga et al., 2019), applying mbed computation for guide tree generation (Blackshields et al., 2010). The command structure was:

```
t_coffee -reg -seq [input].fasta -nseq 100 -tree mbed \  
-method mafftginsi_msa -outfile [output]_aln.fasta \  
-outtree [output].mbed -thread 9
```

MAFFT Multiple Alignment using Fast Fourier Transform (MAFFT) G-INS-i progressive algorithm (Kato and Standley, 2013) was executed either directly or through T-COFFEE. For region-specific realignment, the Geneious implementation of MAFFT version 7.490 (Kato and Standley, 2013) was used with algorithm automatically determined, BLOSUM80 scoring matrix, gap penalty of 1.7, and offset value of 0.123.

2.3.2 Alignment optimisation

Alignment approaches were systematically compared by constructing pilot phylogenies and evaluating:

- Statistical support for topology (ultrafast bootstrap values)
- Concordance with the One Thousand Plant Transcriptomes Initiative species tree (Leebens-Mack et al., 2019)

Manual alignment curation (Morrison, 2006) was tested at varying stringency levels using Geneious 2020.0.2 (Biomatters, New Zealand). Region-specific alignment was used to re-align low-signal regions before manual curation, accounting for rate heterogeneity across different protein regions (Bevan et al., 2007).

2.3.3 Profile alignments

Profile alignments combining sequences from different clusters were built using T-COFFEE to preserve within-cluster alignment quality while optimising between-cluster alignment:

```
t_coffee [cluster1].fasta, [cluster2].fasta \  
-profile=[sequence files] -output=fasta \  
-outfile=[date]_profile.fasta -profile_comparison=full -thread=6
```

2.3.4 Manual curation

Alignments were manually curated in Geneious 2020.0.2 to remove:

- Truncated sequences lacking an N-terminal methionine that were markedly shorter than their most similar neighbours
- Sequences with significant deletions in highly conserved regions or large deletions likely representing splice variants or pseudogenes
- Sequences with undefined residues (X characters)
- Duplicate sequences arising from database redundancy

2.3.5 Transit peptide removal

Putative N-terminal transit peptides were removed prior to phylogenetic analysis as these sequences are cleaved post-translationally and do not contribute phylogenetic signal relevant to the mature functional protein. Transit peptides evolve under different selective pressures than the mature protein and can introduce alignment artefacts.

For plant nucleotide phosphatases, AtSAL1 has an experimentally determined transit peptide with a well-conserved cleavage site at a 'MAYE' amino acid motif (Estavillo et al., 2011). Alignments were manually curated to ensure this motif was correctly aligned across all sequences. Using the AtSAL1 cleavage site as reference, all residues upstream of this motif were removed. For sequences where the MAYE motif was not conserved, the corresponding alignment position was used as the cleavage point.

2.3.6 Case-study specific parameters

Phytoene synthase

The Multiple Sequence Alignment (MSA) was built using T-COFFEE version 13.45.61.3c310a9 (Notredame et al., 2000) with default parameters. The alignment was visualised in Geneious Prime version 2022.0.2. As the signal peptide cleavage site is well conserved, putative signal peptides were removed at the cleavage site of *A. thaliana* PSY.

The alignment was manually curated where duplicate or paralogous proteins were removed, as well as those with large deletions or insertions (defined

as >5% of the mature *A. thaliana* PSY sequence length). Where multiple sequence copies existed for the same species, the copy bearing highest similarity to *A. thaliana* PSY was retained. The poorly conserved N-terminal loop and C-terminus were realigned using the Geneious implementation of MAFFT version 7.490 (Kato and Standley, 2013) with algorithm automatically determined, BLOSUM80 scoring matrix, gap penalty of 1.7, and offset value of 0.123. The final MSA comprised 116 sequences.

Decapping exoribonuclease

Sequence alignments were performed with T-COFFEE using default settings (Notredame, 2010).

Wheat SAL

Sequences were aligned using T-COFFEE Version 13.45.61.3c310a9 (Notredame et al., 2000) with default parameters. The MSA was visualised in Geneious Prime version 2022.0.2. Regions requiring further alignment were realigned using the Geneious implementation of MAFFT version 7.490 (Kato and Standley, 2013) with algorithm automatically determined, BLOSUM80 scoring matrix, gap penalty of 1.7, and offset value of 0.123, followed by manual curation. Sequences with large insertions or deletions inconsistent with the consensus were removed as possible splice variants.

2.4 Phylogenetic inference

2.4.1 Evolutionary model selection

Phylogenies were constructed using IQ-TREE2 (Minh et al., 2020). The optimal evolutionary model was identified using ModelFinder (Kalyaanamoorthy et al., 2017), testing nuclear-encoded amino acid substitution models with both the FreeRate and Gamma rate heterogeneity models. The model search was restricted to empirical amino acid substitution matrices appropriate for nuclear-encoded proteins.

Initial model testing used the command:

```
iqtree2 -s [alignment].fasta -m MF \
      -mset WAG, LG, JTT, UL2, UL3, EX_EH0, LG4M, LG4X, CF4 -T AUTO
```

This tested the following model classes:

- Standard empirical matrices: WAG, LG, JTT
- Protein mixture models: UL2, UL3 (Le et al., 2008), LG4M, LG4X (Le and Gascuel, 2010)
- Extended empirical models: EX_EHO, CF4

Model selection was based on both the Akaike Information Criterion (AIC) (Akaike, 1974) and Bayesian Information Criterion (BIC) (Schwarz, 1978).

2.4.2 Maximum likelihood tree inference

Test phylogenies were initially constructed with 1,000 maximum likelihood tree searches and 1,000 ultrafast bootstrap replicates per tree (Minh et al., 2013; Hoang et al., 2018):

```
iqtree2 -s [alignment].fasta -m UL3 -B 1000 -T AUTO
```

For final analyses using the optimal model, replicate phylogenies were generated on the National Computing Infrastructure (NCI) GADI supercomputer:

```
iqtree2 -s [alignment].fasta -st AA -m [MODEL] \
  -bb 1000 -nt AUTO -ntmax 8
```

Key parameters:

- Sequence type: amino acid (-st AA)
- Ultrafast bootstrap replicates: 1,000 per tree
- Thread allocation: automatic with maximum of 8

2.4.3 Dataset reduction for outgroup clades

To reduce computational burden while maintaining phylogenetic signal, sequences in outgroup clades were reduced using T-COFFEE's reformatting utility. These clades were first profile-aligned:

```
t_coffee -profile=[sequence files] -output=fasta \
  -outfile=[date]_branch_outgroups.fasta \
  -profile_comparison=full -thread 6
```

Then trimmed to retain representative sequences:

```
t_coffee -other_pg seq_reformat -in [input].fasta \
  -action +trim _seq_n100 -output fasta_seq > [output].fasta
```

2.4.4 Statistical validation

Replicate trees were tested for statistical bias using the approximately unbiased (AU) test (Shimodaira, 2002) implemented in IQ-TREE2:

```
iqtree2 -s [alignment].fasta -z all_trees.nwk \
    -te [alignment].fasta.treefile -m [MODEL] \
    -zb 100000 -au -T 6
```

Parameters:

- Number of RELL bootstrap replicates: 100,000 (-zb 100000)
- AU test enabled (-au)

2.4.5 Consensus tree construction and final tree selection

Consensus cladograms were constructed from the highest-scoring trees ranked by log-likelihood (lnL) combined with AU test *p*-values:

```
iqtree2 -sup [input_trees].nwk [output].nwk
```

The consensus topology was qualitatively compared against the One Thousand Plant Transcriptomes Initiative phylogeny (Leebens-Mack et al., 2019) to identify the replicate tree most consistent with established species relationships. The tree from the replicates that best matched this consensus topology was selected as the final phylogeny for downstream analysis.

2.4.6 Tree rooting

For plant nucleotide phosphatases, an initial unrooted phylogeny was constructed, and the closest branch below the SAL clade with greater than 95% ultrafast bootstrap support was selected as the root. This rooting strategy placed out-group sequences (from Chromista, Rhodophyta, and Glaucophyta) appropriately and enabled clear delineation of the three major nucleotide phosphatase clades (SAL, AHL, and CHLOROPLASTIC NUCLEOTIDE PHOSPHATASE-LIKE (CNP)).

2.4.7 Case-study specific parameters

Plant nucleotide phosphatases

The optimal model was EX_EHO+R9 (BIC score 285,583.624), with JTT+F+R9 (BIC score 287,539.190) and LG+R10 (BIC score 287,550.402) as the second

and third best-performing models respectively. Four hundred replicate phylogenies were generated using the optimal model (EX_EHO+R9: extended empirical model with 9 FreeRate categories). All 400 replicate trees passed the AU test, indicating no statistically significant bias in tree topology. A consensus cladogram was constructed from the 20 highest-scoring trees, and the tree most closely matching this consensus topology was selected as the final phylogeny.

Wheat SAL

ModelFinder was used to search against all available protein models. Phylogenetic reconstruction used 10 replicates of each of the two highest-scoring evolutionary models, each with 1,000 ultrafast bootstraps (Hoang et al., 2018). The best phylogeny was selected by manual evaluation, first removing all trees with a non-contiguous outgroup. The phylogeny was visualised as a cladogram using R packages ggtree (Yu et al., 2017) and ape (Paradis and Schliep, 2019).

2.5 Evolutionary velocity analysis

Evolocity (Hie et al., 2022) uses protein language models to convert protein sequences into high-dimensional embeddings that capture evolutionary and functional relationships beyond traditional sequence alignment. The method constructs a vector field of evolutionary change based on local mutational likelihoods, enabling analysis of large sequence datasets where traditional phylogenetic reconstruction becomes computationally impractical.

2.5.1 Network construction and velocity embedding

K-Nearest Neighbours (KNN) networks were constructed from sequence embeddings. Dimensionality reduction was performed using Uniform Manifold Approximate Projection (UMAP) (Healy and McInnes, 2024). Velocity embeddings were generated to visualise evolutionary trajectories.

2.5.2 Case-study specific parameters

Decapping exoribonuclease

Pseudotime analysis was performed using Evolocity, executed with default parameters except nearest-neighbour clustering at 25 sequences.

Eukaryotic translation initiation factor

Evolocity analysis was performed using a **KNN** network with neighbours = 150. Dimensionality reduction used **UMAP**. Velocity embedding parameters were: basis= 'umap', min_mass=0.5, smooth=1.0, linewidth=1.0, density=1.0, size=50.

2.6 Structural modelling

2.6.1 AlphaFold2 and ColabFold

Structural models were built using the ColabFold ([Mirdita et al., 2022](#)) implementation of AlphaFold2 ([Jumper et al., 2021](#)). Structures were inferred from amino acid sequences with signal peptides removed, using the 'no template' method. Predicted structures were relaxed using amber force fields. Other settings were as default. Five structures were predicted and ranked according to the **pLDDT** confidence measure ([Mariani et al., 2011](#)), with the highest-ranked model selected.

2.6.2 ESMFold

ESMFold ([Lin et al., 2023](#)) was used for independent structural validation. Models were generated using one copy and three recycles.

2.6.3 Model validation

Structural models were validated by comparison with experimentally determined crystal structures of homologous proteins. Structural alignments were made using PyMOL ([Schrödinger, 2020](#)). Figures were generated using PyMOL.

2.6.4 Case-study specific parameters

Phytoene synthase

The ColabFold model achieved a **predicted local distance difference test (pLDDT)** score of 88. The ESMFold model achieved a **pLDDT** score of 87. While the ESMFold model had higher **pLDDT** scores for the N- and C-termini and the N-terminal variable loop region, the **pLDDT** scores for the conserved core of the enzyme were marginally lower. Therefore, the ColabFold model was chosen.

The fold was validated by comparing the model with the crystal structure of *Enterococcus hirae* dehydrosqualene synthase complexed with three Mg^{2+} ions and the farnesyl thiopyrophosphate-FPS ligand (PDB accession: 5IYS) (Figure 3.2).

2.7 Sequence logo analysis

Sequence logos were generated using WebLogo (Crooks et al., 2004) to visualise conservation at specific positions. For PSY analysis, logos showed four residues each side of mutation sites across the final MSA of 116 sequences spanning gymnosperms and angiosperms.

2.8 Functional annotation

2.8.1 Subcellular localisation prediction

Subcellular localisation was predicted using the deep learning algorithm DeepLoc 2.0 (Armenteros et al., 2019), which predicts localisation to 10 subcellular compartments based on protein sequence. As DeepLoc 2.0 showed limited sensitivity for mitochondrial targeting in preliminary tests (failing to predict known mitochondrial localisation for AtSAL1), MitoFates (Fukasawa et al., 2015) was additionally used to specifically assess mitochondrial targeting potential. MitoFates detects cleavable N-terminal presequences characteristic of mitochondrial matrix proteins.

Putative transit peptides identified by these tools were manually reviewed against the AtSAL1 experimentally determined transit peptide and the mature protein N-terminus.

2.8.2 Co-expression analysis

Co-expression analysis was performed using ATTED-II (<https://atted.jp/>) (Obayashi et al., 2018). For *A. thaliana* nucleotide phosphatases (AtSAL1, AtAHL, and AtCNP), genes with the highest support score (3) and Mutual Rank score greater than 3 (the default ATTED-II threshold for high-confidence co-expression) from the 'auth-u.2' dataset were extracted for gene ontology analysis.

2.8.3 Gene ontology enrichment

Gene ontology enrichment analysis was performed using PANTHER (<http://www.pantherdb.org/>) (Thomas et al., 2003). Co-expressed gene lists were analysed using the statistical over-representation test with binomial distribution and false discovery rate correction. Three ontology categories were tested:

- GO Biological Process
- GO Molecular Function
- GO Cellular Component

Results were filtered to retain terms with p -value $< 1 \times 10^{-4}$, false discovery rate $< 1 \times 10^{-2}$, and fold enrichment > 30 . Where more than three terms passed these thresholds, only the top three by fold enrichment were reported.

2.8.4 Tissue and condition-specific expression

Tissue-specific and condition-responsive expression patterns were examined using Genevestigator 8.3.2 (NEBION AG, Switzerland) for *A. thaliana* Col-0 ecotype. Expression data for AtSAL1, AtAHL, and AtCNP were extracted from experiments showing significant expression changes between control and treated conditions, including responses to pathogen infection (*Phytophthora infestans*, *Pseudomonas syringae*) and severe drought stress (10 days without water).

2.9 Statistical methods

2.9.1 Comparison of sequence properties between clades

To assess functional divergence between putative outparalogous clades, statistical comparisons of sequence properties were performed. Three properties were compared:

1. Sequence length (amino acids)
2. Aromaticity (proportion of aromatic residues: Phe, Trp, Tyr)
3. Isoelectric point (theoretical pI)

These properties were calculated using BioPython's ProteinAnalysis module (Cock et al., 2009). Comparisons between clades were performed using Welch's t -test, which does not assume equal variances between groups. Effect sizes were calculated using Cohen's d and interpreted as:

- negligible ($|d| < 0.2$),
- small ($0.2 \leq |d| < 0.5$),
- moderate ($0.5 \leq |d| < 0.8$), or

- large ($|d| \geq 0.8$).

Statistical analyses were performed in R 4.0.2 (R Core Team, 2019) using the `rstatix` package (Kassambara, 2020). Significance thresholds were set at $p < 0.05$, with additional thresholds at $p < 0.01$, $p < 0.001$, and $p < 0.0001$.

2.9.2 Conservation of translation initiation factor isoforms

To test whether `eIF(ISO)4E` is conserved across plant lineages, the distribution of `eIF(ISO)4E` homologues was compared between non-angiosperm ($n = 12$ orders) and angiosperm ($n = 31$ orders) plant groups. `EIF4E`, found in all 43 orders, served as a methodological control. The null hypothesis that `eIF(ISO)4E` is conserved across all plants was tested against the alternative that `eIF(ISO)4E` presence is reduced in non-angiosperms.

Non-parametric tests were employed to account for zero values in the dataset: one-tailed Mann-Whitney U test on `eIF(ISO)4E` counts and one-tailed Fisher's exact test on presence/absence data. Significance was assessed at $\alpha = 0.05$. Power analysis used observed effect sizes (Cohen's d , ϕ coefficient) with statistical power ≥ 0.80 . Analyses assumed phylogenetic independence among orders.

To test whether `eIF(ISO)4E` and `eIF4E` differ in count number within angiosperm orders, paired comparison was performed across angiosperm ($n = 31$) orders using the Wilcoxon signed-rank test with Pratt's method for handling zeros. A two-tailed alternative hypothesis tested whether the median difference between `eIF4E` and `eIF(ISO)4E` counts is zero. Statistical significance was assessed at $\alpha = 0.05$. Statistical tests were performed using Python with `scipy.stats`.

2.9.3 Serine-rich domain identification

For `DXO1` analysis, the presence of a serine-rich N-terminal domain was determined by calculating the frequency of serine residues within the first 18% of amino acids for each homologue, based on manual curation of sequence alignments. A threshold frequency of 0.23 was determined based on the trough in its bimodal distribution, above which was considered to reflect the presence of this region (Figure 3.4C). Sequences with fewer than 10 serines were excluded to avoid artefacts from short sequences.

Cladograms were based on hierarchical relationships between taxa defined in `Pfam`, with the following deviations:

- Zygnematophyceae was placed as sister group to Embryophyta,
- mosses and liverworts were grouped together, and

- Amborellales was moved to be sister to all other angiosperms, followed by Nymphaeales.

2.10 Computational tools and software

Table 2.2 summarises the computational tools used for sequence-based classification and their primary applications.

Table 2.2: Computational tools used for sequence-based classification

Tool	Application	Citation
BLAST/TBLASTN	Sequence database searching	Altschul et al. (1990)
CD-HIT	Sequence clustering and redundancy reduction	Li and Godzik (2006); Huang et al. (2010)
EFI-EST	Sequence similarity network generation	Gerlt et al. (2015)
Cytoscape 3.7.2–3.9.1	Network visualisation and analysis	Shannon et al. (2003)
T-COFFEE 13.45.61.3c310a9	Multiple sequence alignment	Notredame (2010)
MAFFT 7.490	Multiple sequence alignment	Katoh and Standley (2013)
Geneious Prime 2020.0.2	Manual alignment curation and visualisation	Biomatters (New Zealand)
IQ-TREE2	Maximum likelihood phylogeny inference	Minh et al. (2020)
ModelFinder	Evolutionary model selection	Kalyaanamoorthy et al. (2017)
Evolocity	Evolutionary velocity analysis	Hie et al. (2022)
ColabFold	AlphaFold2 structural modelling	Mirdita et al. (2022)
ESMFold	Structural modelling	Lin et al. (2023)
BioPython	Sequence processing and analysis	Cock et al. (2009)
BioServices	Programmatic database access	Cokelaer et al. (2013)
DeepLoc 2.0	Subcellular localisation prediction	Armenteros et al. (2019)
MitoFates	Mitochondrial targeting prediction	Fukasawa et al. (2015)
ATTED-II	Gene co-expression analysis	Obayashi et al. (2018)
PANTHER	Gene ontology enrichment	Thomas et al. (2003)
Genevestigator 8.3.2	Expression analysis	NEBION AG (Switzerland)
WebLogo	Sequence logo generation	Crooks et al. (2004)
PyMOL	Structural visualisation and alignment	Schrödinger (2020)
R 4.0.2	Statistical analysis and visualisation	R Core Team (2019)

2.10.1 Code and data availability

Custom scripts for sequence processing, filtering, and analysis were written in Python 3 (Jupyter Lab notebooks) and R (RStudio notebooks). Code is available at <https://github.com/mattmorts-sci/>. Key packages used include:

- **Python:** pandas (McKinney, 2010), numpy (Harris et al., 2020), BioPython (Cock et al., 2009), matplotlib (Hunter, 2007), seaborn (Waskom, 2021), scipy.stats
- **R:** tidyverse (Wickham et al., 2019), ggtree (Yu et al., 2017), treeio (Wang et al., 2020), ape (Paradis and Schliep, 2019), rstatix (Kassambara, 2020)

Chapter 3

Sequence and structural approaches to protein classification

Chapter outline

This chapter covers sequence classification work for three separate projects:

- Phytoene synthase
- Decapping exoribonuclease (DXO1)
- Eukaryotic translation initiation factors

The Introduction and Discussion cover all three of these projects, while the Results section is broken down into three subsection – one for each of the projects.

Publishing Contributions

This chapter presents computational analyses contributed to three collaborative research projects.

Phytoene synthase: As a middle author on this paper, I performed all data acquisition, cleaning and curating for the sequence logo and structural modelling. In particular, I contributed the structural modelling and analysis to identify the mechanistic basis for reduced **PSY** function in these mutants. I wrote the relevant sections in the methods and results, and contributed to the discussion. This work was published in [Hou et al. \(2024\)](#).

Decapping exoribonuclease: This paper has been submitted to Proceedings of the National Academy of Sciences of the United States of America (PNAS) and is available in bioRxiv (<https://doi.org/10.1101/2025.10.29.684918>). I contributed all of the evolutionary and classification research which has been repro-

duced below. I performed all data acquisition, cleaning and curating, and analysis for the evolutionary context and subsequent classification of plant-DXO1. I wrote the evolutionary section of the paper, as well as the relevant methods section. I contributed to the design and underpinning intellectual thinking of the paper.

Eukaryotic translation initiation factors: This paper is in draft and I will be a middle author. Building on the success of the DXO1 analysis, I was asked to establish the evolutionary trajectory of eIF4E, and determine whether our collaborator's earlier work (Patrick and Browning, 2012) is still supported given that at the time of their study they had only a fraction of the sequences now available.

Significance statements

Phytoene synthase

- **Key Discovery:** Four mutations in *Arabidopsis* PSY reduce enzyme activity through distinct mechanisms: disrupting substrate binding, protein-protein interactions with ORANGE protein (OR) and GGPP SYNTHASE 11 (GGPPS11), and metal cofactor coordination.
- **Methodological Approach:** Structural modelling using AF2 and ESMFold revealed how specific amino acid substitutions in PSY affect enzyme structure and function, providing mechanistic insights into reduced carotenoid flux.
- **Biological Significance:** PSY mutations modulate cis-carotene accumulation, reducing carotenoid-derived signalling that causes phenotype rescue in carotenoid and chloroplast regulation 2 (*ccr2*) plants without causing lethality.

Decapping exoribonuclease

- **Methodological Approach:** The Evolocity network analysis tool overcomes the computational limits of traditional phylogenetics for large datasets (>500 sequences), providing an efficient and robust method for evolutionary analysis and classifying gene forms.
- **Key Evolutionary Discovery:** Angiosperms evolved a unique form of DXO1 characterised by a novel serine-rich N-terminal domain, a key functional innovation that arose early in angiosperm evolution.

- **Biological Significance:** The ancestor of modern [DXO1](#) underwent neo-functionalisation, where a new plant-specific domain acquired regulatory control over mRNA surveillance and ribosome loading, resulting in a specialised adaptation in plant RNA biology.

Eukaryotic translation initiation factors

- **Methodological Approach:** The [Evolocity](#) network analysis tool, combined with sequence similarity networks and formal statistical testing, provided an efficient method for determining evolutionary directionality and classifying gene isoforms.
- **Key Evolutionary Discovery:** Angiosperms evolved a unique form of the translation initiation factor [eIF\(ISO\)4E](#) through gene duplication of [eIF4E](#), confirming with high statistical confidence what was previously uncertain due to limited sampling.
- **Biological Significance:** [EIF\(ISO\)4E](#) underwent neo-functionalisation following duplication, acquiring specialised regulatory control over specific mRNA subsets and localised translation, representing an essential adaptation in angiosperm translational control.

3.1 Introduction

Classifying proteins into functional groups is fundamental to understanding biological systems. Homologous proteins that share evolutionary ancestry may diverge in function, while convergent evolution can produce functional similarity without shared ancestry. Computational approaches therefore have the challenge of identifying functional similarity that may not be congruent with evolutionary signal. This challenge is more acute when classifying paralogues and variants that diverged recently or differ subtly.

Two broad computational strategies address this challenge. Structural approaches predict or compare three-dimensional protein architecture, enabling identification of functionally important sites even when sequence identity is low. Static structural models, whether experimentally determined or computationally predicted, reveal active site geometry, substrate binding pockets, and mutation effects on protein stability. Structural similarity has not been feasible until recently due to the lack of experimentally determined structures or structural models. Recent advances in structural modelling have made this approach more feasible. Sequence-based approaches exploit the correlation between sequence similarity and functional conservation, using alignment

scores, clustering thresholds, or phylogenetic reconstruction to group proteins. Sequence similarity methods scale efficiently to large datasets but assume that sequence similarity reliably predicts function.

This chapter presents three case studies that demonstrate the application of these computational approaches to protein classification problems in plants. The first uses static structural modelling to understand how point mutations affect enzyme function. The second and third use sequence-based evolutionary analysis to classify protein variants and test evolutionary hypotheses. Brief overviews of *SSNs*, *Evolocity* and phylogenetics can be found in chapter 1 1.2.2. Each case study reveals characteristic limitations that motivate the methodology developed in subsequent chapters.

3.1.1 Phytoene synthase

PSY catalyses the first committed step in the carotenoid biosynthesis pathway. It catalyses the synthesis of 15-*cis*-phytoene (phytoene) from two *GGPP* molecules which is then passed further down the pathway to ultimately produce various carotenoids (Zhou et al., 2022). Carotenoids are a diverse group of hydrophobic isoprenoid compounds that provide pigment and aroma almost exclusively in plants, serving roles in photoprotection, photosynthesis, and as precursors for signalling molecules (Baranski and Cazzonelli, 2016; Cazzonelli, 2011).

PSY expression is regulated by transcription factors that respond to light signals. *PHYTOCHROME INTERACTING FACTOR 3* (*PIF3*) suppresses *PSY* expression in darkness, while *ELONGATED HYPOCOTYL 5* (*HY5*) promotes its expression in light (Hou et al., 2024). The balance between *PIF3* and *HY5* coordinates carotenoid biosynthesis with light availability during seedling development. High *PIF3*/*HY5* ratios maintain dark development (skotomorphogenesis), while low ratios promote light-induced development (photomorphogenesis). When this regulatory balance is disrupted, plants develop abnormal plastid structures and pigmentation defects.

PSY's activity is post-transcriptionally regulated by *OR* through protein-protein interactions (Zhou et al., 2015). The *OR* protein stabilises *PSY* and prevents its degradation, thereby increasing carotenoid accumulation. Several methods are used to study these interactions: yeast two-hybrid assays detect protein binding in living yeast cells, while co-immunoprecipitation captures protein complexes from plant extracts. Additionally, *GGPPS11* has been demonstrated as essential for *PSY* phytoene production, forming a multi-enzyme complex that channels substrate to *PSY* (Ruiz-Sola et al., 2016). This complex form-

ation is necessary for efficient carotenoid biosynthesis *in planta*.

The *A. thaliana ccr2* mutant has a loss of function in CAROTENOID ISOMERASE (CRTISO), causing accumulation of linear cis-carotenes and developmental abnormalities including leaf virescence, impaired chloroplast biogenesis, and abnormal photomorphogenic responses (Hou et al., 2024). Virescence refers to the yellowing of normally green tissues – in *ccr2*, new leaves emerge yellow and gradually green as they mature, indicating delayed chloroplast development. These phenotypes result from accumulation of a carotenoid-derived signal that disrupts normal plastid development signalling pathways.

In our study (Hou et al., 2024), we identified four EMS-induced mutations in PSY that rescue the *ccr2* phenotypes: *psy-4* (M266I), *psy-90* (A352T), *psy-130* (P178S), and *psy-145* (splice site mutation). *Psy-145* will not be discussed in this section as the mutation was not explicitly structural. These mutations reduce flux through the carotenoid pathway without causing complete lethality, providing a unique opportunity to understand PSY structure-function relationships.

3.1.2 Decapping exoribonuclease

DXO1 is the plant homologue of the conserved eukaryotic Rai1/Dxo1/DXO enzyme family, which controls RNA quality at the 5'-end. *A. thaliana* DXO1 retains the family's core NAD⁺ cap hydrolase (deNADding) activity but has evolved distinct features that set it apart from fungal and mammalian counterparts. It lacks the domain needed for XRN interaction in plants and contains amino acid substitutions that specialise its activity towards deNADding rather than general decapping or exoribonuclease functions (Xiao et al., 2023; Jiao et al., 2013; Pan et al., 2020; Chang et al., 2012; Doamekpor et al., 2020a,b). Most significantly, it contains a plant-specific, intrinsically disordered serine-rich N-terminal domain that is essential for plant development (Kwasnik et al., 2019; Pan et al., 2020).

DXO1 functions through both deNADding-dependent and independent mechanisms. The deNADding activity removes NAD⁺ caps from target RNAs, particularly ribosomal protein transcripts, marking them for degradation (Yu et al., 2021). The serine-rich N-terminal domain provides an additional, deNADding-independent, function by enabling nuclear interaction with RNMT1. This then activates m7G cap methylation (Xiao et al., 2023). This DXO1-RNMT1 partnership is a plant-specific adaptation that parallels the mammalian RN-MT/RAM system (Xiao et al., 2023; Gonatopoulos-Pournatzis et al., 2011).

Loss of DXO1 causes severe developmental defects including stunted growth,

malformed chloroplasts, and floral abnormalities. These phenotypes arise specifically from loss of the serine-rich domain rather than deNADding activity (Kwasnik et al., 2019; Pan et al., 2020; Zakrzewska-Placzek et al., 2025). Deleting this domain causes mis-expression of ribosomal genes and accumulation of pre-rRNAs, resulting in ribosomal abnormalities (Zakrzewska-Placzek et al., 2025; Xiao et al., 2023). These findings demonstrate that DXO1 has evolved beyond its ancestral RNA surveillance role to become essential in many plants for plant-specific mRNA regulation and development. The serine-rich N-terminal domain represents a key evolutionary innovation but which had not yet had the evolutionary origin determined. Classifying plant DXO1 is therefore complicated based on the presence or absence of this domain.

3.1.3 Eukaryotic translation initiation factors

The translation initiation factor eIF4E serves as the primary translation factor for essential developmental processes (Liu et al., 2022; Zhang et al., 2024), structured mRNAs (Mayberry et al., 2009), and basal stress responses (Salazar-Díaz et al., 2021). The plant-specific isoform eIF(ISO)4E adds regulatory flexibility (Patrick and Browning, 2012), enabling fine-tuned control over specific mRNA subsets (Mayberry et al., 2009; Salazar-Díaz et al., 2021) and localised translation (Patrick and Browning, 2012).

Both translation initiation factors eIF4E and eIF(ISO)4E show broad expression across angiosperm plant tissues (Liu et al., 2022; Zimmermann et al., 2004; Winter et al., 2007). Patrick and Browning (2012) proposed that eIF(ISO)4E arose in angiosperms after identifying it in *Amborella trichopoda* but not in five non-flowering plant genomes in their dataset.

However, such a small sample size cannot definitively determine whether the absence of eIF(ISO)4E reflects evolutionary divergence in angiosperms or insufficient sampling. Testing the null hypothesis that eIF(ISO)4E is conserved across all plant lineages requires genomic data from representative species across all plant groups. As sampling increases across diverse taxonomic groups, we gain greater statistical power and confidence about conservation patterns.

The significant expansion of genomic and transcriptomic data since the study of Patrick and Browning (2012) now allows us to test their assertion more rigorously. I also sought to evaluate alternative evolutionary mechanisms such as convergent evolution or lateral gene transfer. Lateral gene transfer is difficult to prove (Bock, 2010) and should not be the null hypothesis, while convergent evolution at the molecular level involves sequences from evolutionarily distant organisms becoming more similar after initial divergence. Building on the suc-

cess of using **Evolocity** for **DXO1** classification (Section 3.2.2), I applied this method to the **eIF4E** family. **Evolocity** uses protein language models to generate a **KNN** sequence map and predict evolutionary directionality by constructing a vector field of evolutionary change based on local mutational likelihoods, enabling insights from large sequence datasets where traditional phylogenetic reconstruction becomes computationally impractical.

The following sections present results from each case study in turn, beginning with structural analysis of **PSY** mutations, then moving to sequence-based evolutionary analyses of **DXO1** and finally **eIF4E/eIF(ISO)4E**.

3.2 Results

3.2.1 Pytoene synthase

I sought to identify the mechanisms of reduced function and whether these mechanisms are likely transferable across modern plants. This was done by building an AF2 model of PSY and examining the sites of the mutations and analysing the likely impacts of the specific mutations. Gymnosperms and angiosperms sequence were aligned and sequence logos created to assess conservation of mutated residues. Sequence logo analysis was conducted to assess conservation of the mutated residues across plant species. Analysis of 116 PSY sequences spanning gymnosperms and angiosperms revealed that all three mutation sites (M266, A352, P178) were fully conserved (Figure 3.1). This complete conservation suggests these residues are critical for the evolved function and that these findings from *A. thaliana* are likely broadly applicable to Viridiplantae.

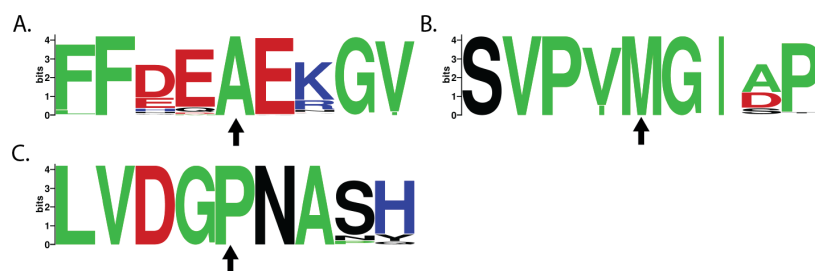


Figure 3.1: PSY mutation site conservation analysis.

Sequence logos showing conservation at mutation sites across 116 plant PSY sequences. Arrows indicate mutation positions. (A) Position 282 (A352T in psy-90) shows complete conservation of alanine. (B) Position 196 (M166I in psy-4) shows complete conservation of methionine. (C) Position 108 (P178S in psy-130) shows complete conservation of proline. Full conservation across Gymnosperms and Angiosperms indicates functional importance of these residues.

Structural models were generated using ColabFold (AF2 implementation) (Mirdita et al., 2022), achieving a pLDDT score of 88. Model validation against *Enterococcus hirae* dehydrosqualene synthase (PDB: 5IYS) confirmed the characteristic ISOPRENOID SYNTHASE fold (Figure 3.2A.) ESMFold (Lin et al., 2023) independently validated the structure (pLDDT: 87) (Figure 3.2B).

The model revealed two conserved aspartate-rich domains essential for catalysis denoted as D172ELVD176 and D298DVGED302. These domains coordinate metal ions required for substrate binding and catalysis, consistent with other isoprenoid synthases (Christianson, 2017).

Analysis of mutation positions revealed three distinct mechanisms of functional impairment (Figure 3.3):

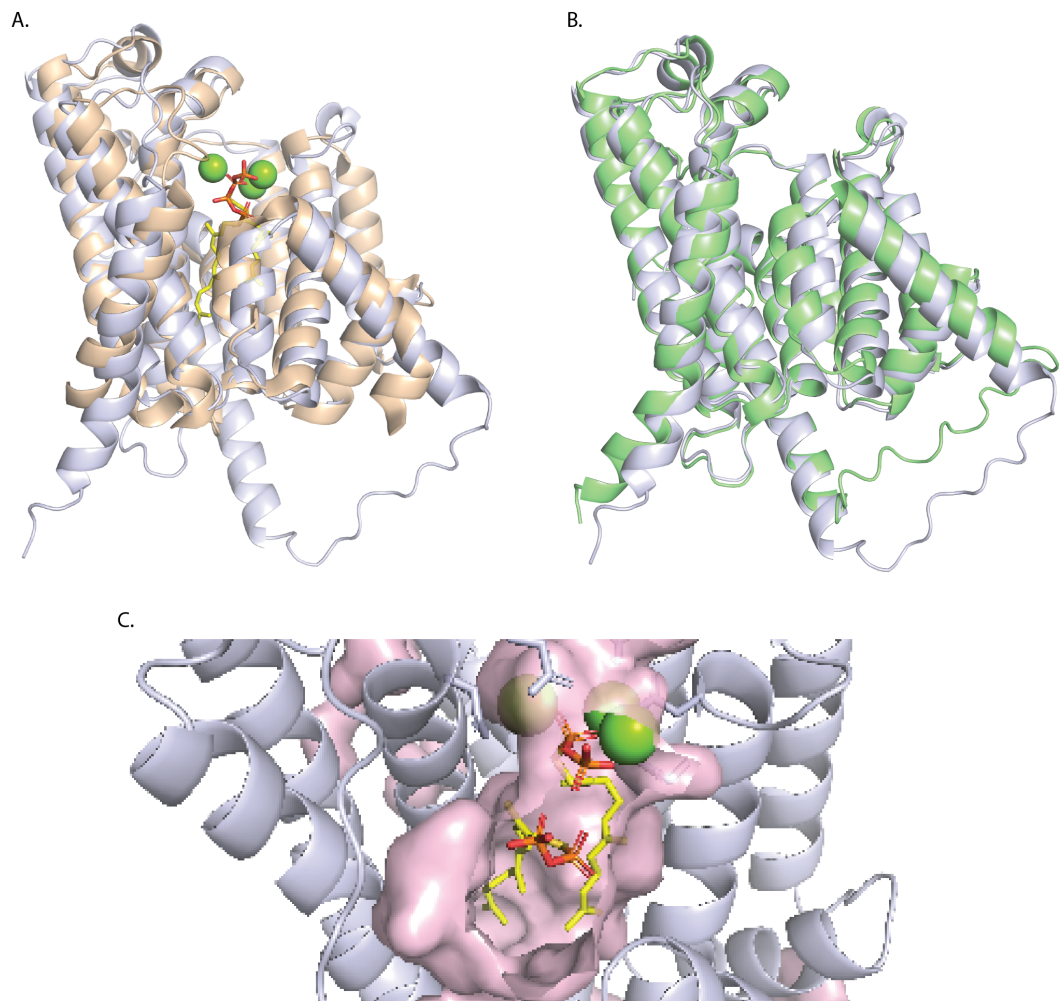


Figure 3.2: PSY structural model validation.

Structural alignment of *Arabidopsis* PSY models with crystal structure homologue. (A) ColabFold model (silver) aligned with *E. hirae* dehydrosqualene synthase (PDB: 5IYS, tan) complexed with Mg^{2+} ions (green spheres) and farnesyl thiopyrophosphate substrate analogue (yellow/orange sticks). (B) ColabFold model (silver) aligned with ESMFold model (green) showing high structural concordance. (C) Substrate binding pocket of ColabFold model with aligned ligands from 5IYS showing conservation of active site architecture.

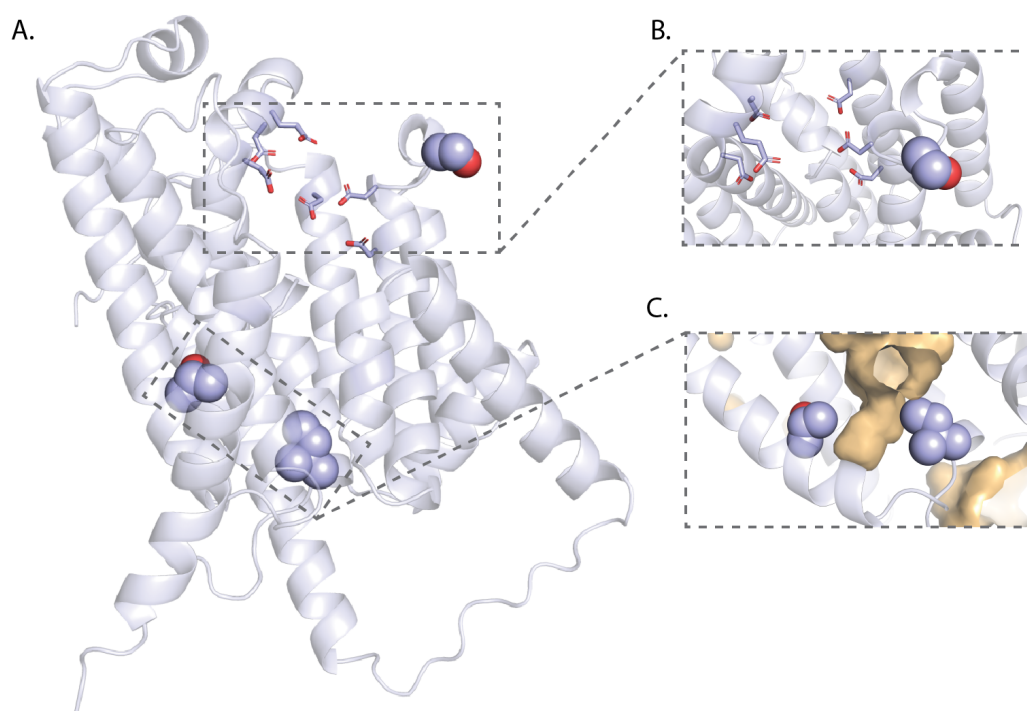


Figure 3.3: Structural mapping of PSY mutations.

Three-dimensional structural model of *Arabidopsis* PSY generated using ColabFold (AF2). Protein backbone shown in cartoon representation with substrate binding pocket (tan) visualised using PyMOL cavity detection (radius 7 Å, cutoff 3 solvent radii). (A) Overview showing mutation positions (spheres) and metal-coordinating aspartate residues (sticks). (B) P178S mutation potentially impacts metal coordination by D172, D176 (same hairpin), and spatially proximal D302. (C) M266I mutation adjacent to substrate pocket base may sterically hinder substrate binding. A352T mutation may disrupt α -helix packing and hydrophobic core stability.

- The M266I substitution is positioned adjacent to the substrate binding pocket. Isoleucine's bulkier side chain when compared to methionine likely causes steric clashes within the pocket, potentially impacting substrate positioning or product release. This variant also showed reduced interaction with OR as demonstrated by yeast two-hybrid and co-immunoprecipitation analyses (Hou et al., 2024).
- The A352T substitution introduces a polar threonine residue into a hydrophobic region proximal to an α -helix. This substitution seems to cause steric clashes with the adjacent α -helix and may disrupt tertiary structure stability through the introduction of a charged moiety into the hydrophobic core. This mutant shows the most severe defects, with near-complete loss of interaction with both OR and GGPPS11, and dramatically reduced phytoene production.
- The P178S substitution replaces a conformationally rigid proline with a flexible serine near the D172ELVD176 metal-binding domain. This increased flexibility could affect the positioning of metal-coordinating aspartate residues, particularly D302 which is spatially proximal to position 178. Despite reduced enzymatic activity, this variant maintains normal interactions with both OR and GGPPS11.

The structural modelling approach identified distinct mutation mechanisms in PSY, but broader evolutionary questions about protein family classification require different methods as raised in the chapter Discussion 3.3.4, and discuss the technical (3.3.1) and biological (3.3.2) insights from this analysis.

3.2.2 Decapping exoribonuclease

This case study addresses how to classify plant DXO1 forms that differ in the presence of a functionally important domain, using sequence-based evolutionary analysis rather than structural modelling. Using Evolocuity I identified the evolutionary trajectory of plant DXO1 using 5,856 high-confidence, non-redundant DXO1 homologue sequences from Pfam Family PF08652, which contains AtDXO1 (AT4G17620.1). The family was identified using Hmmscan (Potter et al., 2018). To understand the dataset and inform further analysis, I first examined basic sequence attributes. Figure 3.4A shows that plant homologues typically have sequence lengths of 500 ± 100 amino acids (aa), longer than fungal homologues, while metazoan sequences show a bimodal distribution. This approximately 200 aa difference matches the length of the serine-rich N-terminal sequence identified in *A. thaliana* (Xiao et al., 2023).

To define the serine-rich N-terminal domain, I analysed serine frequency

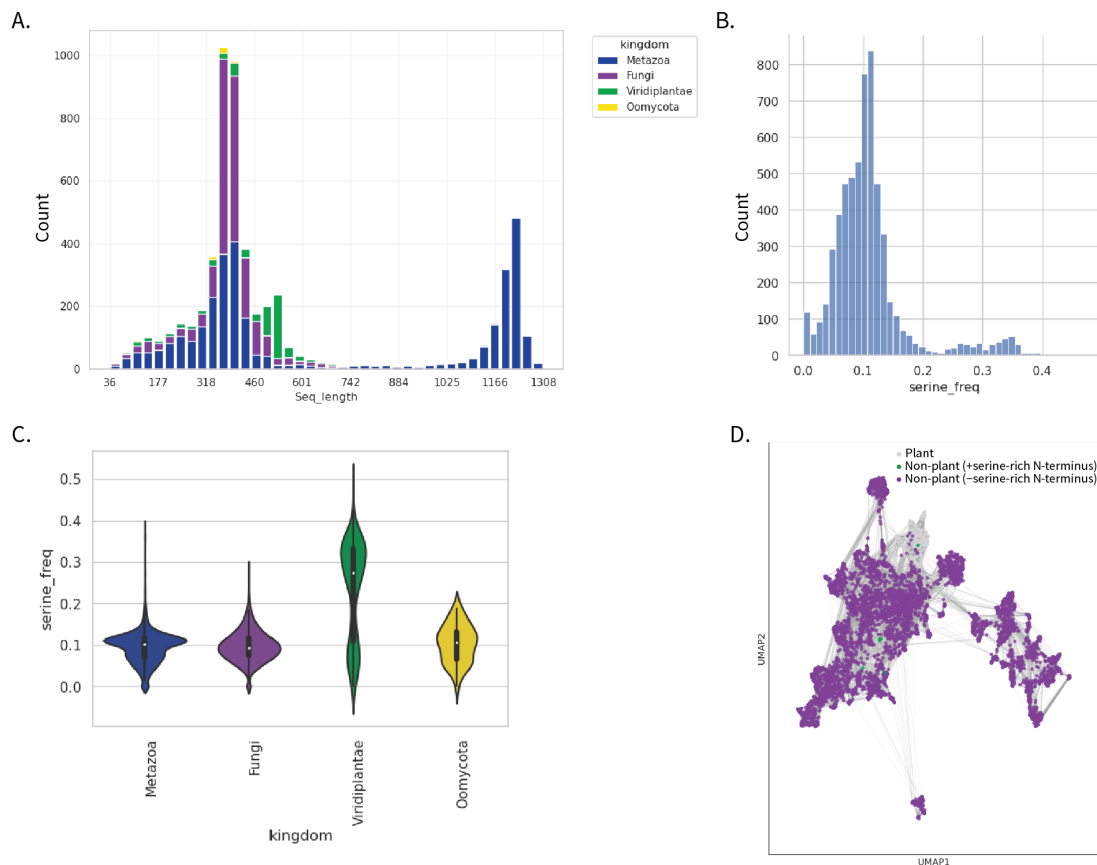


Figure 3.4: Sequence information for Rai1/Dxo1/DXO homologues.

(A) Length distribution of curated and filtered sequences from Pfam family PF08652 comprising the dataset used for evolutionary analysis. Bar colour denotes kingdom. (B) Serine frequency distribution within the first 18% of sequences in our dataset. (C) Violin plot of serine frequency distribution within the first 18% of sequences in our dataset grouped by kingdom. (D) KNN network from Figure 3.5C recoloured to highlight non-plant sequences with (green) and without (purple) a serine-rich N-terminus. Plant sequences are shown in grey.

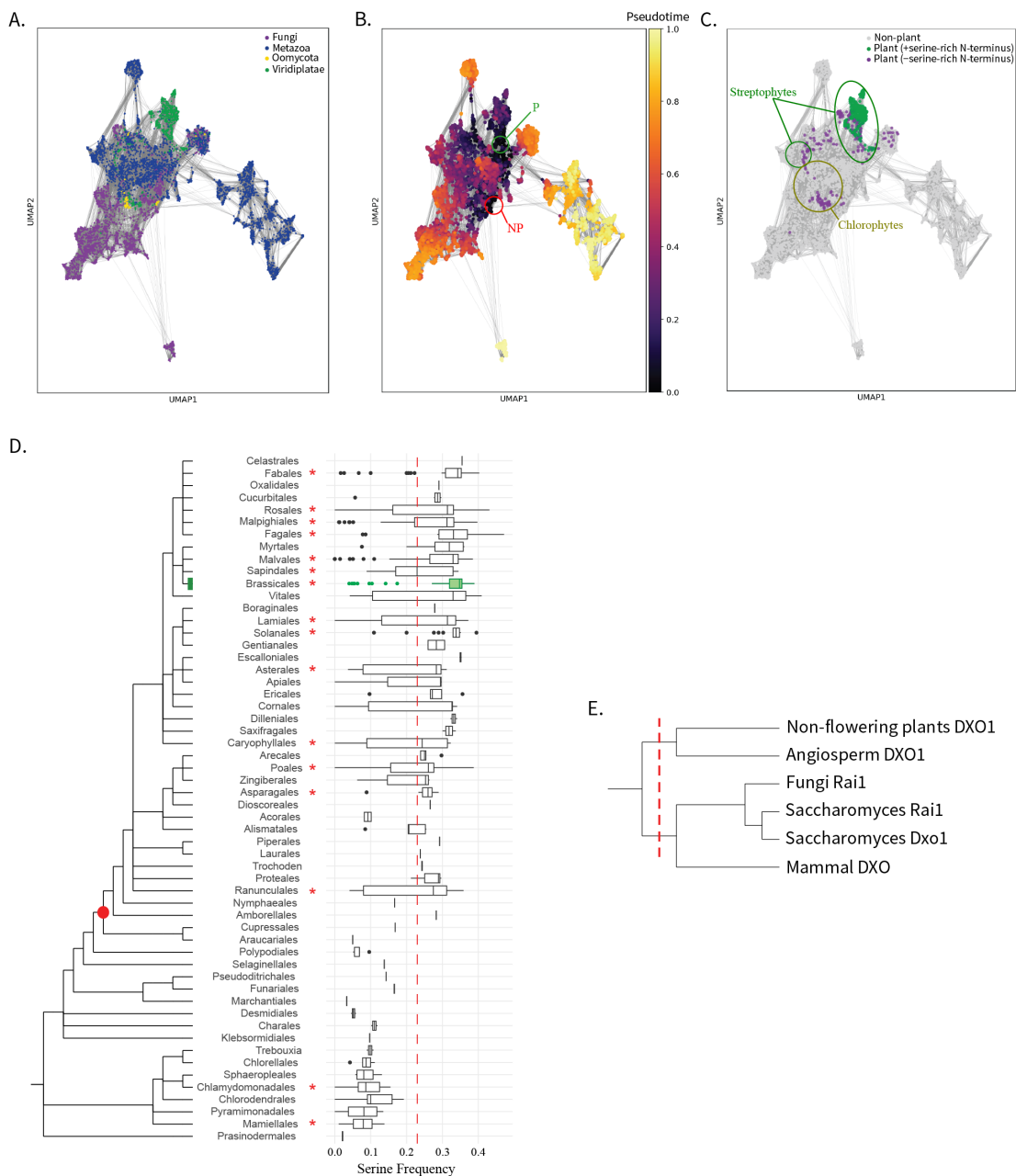


Figure 3.5: Angiosperms harbour a divergent DXO1 protein characterised by a serine-rich N-terminal region.

(A) KNN of DXO1 orthologues using UMAP with 25 nearest neighbours, coloured by taxonomic kingdom. (B) Evolutionary velocity pseudo-time plot with sequences coloured from 0 (ancestral-like) to 1 (most diverged). Two predicted roots marked for plant (P) and non-plant (NP) orthologues. (C) KNN plot from (A) highlighting plant sequences with (green) and without (purple) serine-rich N-terminus. Non-plant sequences in grey. Streptophyte and chlorophyte sequences labelled; unlabelled plant sequences represent mixed clades. (D) Cladogram of plant orders with box-and-whisker plots showing serine frequency. Brassicales highlighted in green. Red node marks angiosperm divergence. Red dotted line indicates serine-rich N-terminus threshold. Serine frequency calculated from first 18% of N-terminal region. Red stars denote clades with $n \geq 10$ sample counts. Full counts are shown in Table 3.1 (E) Order-level cladogram of DXO1 and DXO1-like homologues. Vertical dashed red line indicates cut-off from last common (prokaryotic) ancestor, defining two pseudo-roots.

Table 3.1: Number of DXO1 homologues retrieved per Kingdom

Kingdom	Number of sequences
Bacillariophyta	3
Bigyra	3
Choanoflagellata	1
Chrysophyceae	1
Ciliophora	1
Cryptophyceae	6
Dinophyceae	16
Discosea	2
Endomyxa	3
Euglenozoa	30
Evosea	9
Foraminifera	3
Fungi	1956
Haptophyta	10
Heterolobosea	2
Ichthyosporea	1
Metazoa	3272
Oomycota	54
Pelagophyceae	4
Phaeophyceae	2
Preaxostyla	3
Rhodophyta	19
Tubulinea	1
Viridiplantae	574

Table 3.2: Non-plant DXO1 homologues with serine-rich N-terminals.

ID	Kingdom	Phylum	Class	Order	Freq.
A0A316ZA69	Fungi	Basidiomycota	Exobasidiomycetes	Entylomatales	0.24
A0AAN6GCK9	Fungi	Basidiomycota	Exobasidiomycetes	Tilletiales	0.29
A0AAD5TBT3	Fungi	Chytridiomycota	Chytridiomycetes	Spizello.	0.27
A0A851F7G5	Metazoa	Nematoda	Chromadorea	Rhabditida	0.28
A0AA36BQR8	Metazoa	Mollusca	Cephalopoda	Octopoda	0.36
A0A267DQM7	Metazoa	Platyhelminthes	Rhabditophora	Macrostomida	0.35
A0A267EMI7	Metazoa	Platyhelminthes	Rhabditophora	Macrostomida	0.37
A0A267F4U8	Metazoa	Platyhelminthes	Rhabditophora	Macrostomida	0.36

ID = UniProtKB ID; Freq. = Serine Frequency, Spizello. = Spizellomycetales

in the first 18% (Methods 2.9.3) of each sequence. Figure 3.4B shows an asymmetric bimodal distribution with a trough at a serine frequency of ≥ 0.23 which was chosen as the threshold for 'serine-rich'. I then examined the distribution of serine-rich N-termini across kingdoms (Figure 3.4C) which shows almost exclusive presence in Viridiplantae. This domain occurs likely exclusively in Viridiplantae, with a bimodal distribution indicating distinct groups of plants with high versus low N-terminal serine frequency.

Evolocity pseudo-time analysis identified two likely evolutionary roots (Hie et al., 2022): one within *Fungi* (more similar to metazoan sequences) and another within *Viridiplantae* (Figure 3.5B). No prokaryotic homologues were detected ($E\text{-value} \leq 1$). This dual-root pattern and the absence of a common an-

cestor in the dataset support biochemical evidence that *A. thaliana* DXO1 differs fundamentally from fungal and metazoan homologues (Kwasnik et al., 2019; Pan et al., 2020). Examining plant-only sequences coloured by presence or absence of the serine-rich domain revealed clear patterns (Figure 3.5C). Chlorophytes lack serine-rich N-termini, while most Streptophytes possess them. Streptophytes with serine-rich N-terminals cluster together on the network periphery, while a small group without serine-rich N-terminals clusters within the core metazoan group. Chlorophytes cluster near these Streptophytes that lack the serine-rich N-terminus. The serine-rich N-terminal domain occurs almost exclusively in plants, with only eight exceptions among non-plant sequences (Table 3.2, Figure 3.4D).

To identify evolutionary patterns in the serine-rich N-terminus across plants, I plotted N-terminus serine frequency by plant order, mapped to a cladogram based on Pfam taxonomic hierarchy (Figure 3.5D). This analysis reveals that the serine-rich N-terminus emerged shortly after angiosperm divergence. Some angiosperm clades contain sequences without serine-rich N-termini, indicating multiple gain or loss events.

Figure 3.5E proposes a model of shared lineage within prokaryotes followed by divergence of plant DXO1, then mammalian DXO from fungal Rai1 with the last common ancestor not available in the dataset, as shown by the two pseudo-roots annotated in Figure 3.5B.

The DXO1 analysis demonstrated that Evolocuity can efficiently classify large protein families by evolutionary trajectory. The following case study applies this approach to a different evolutionary question that tests whether eIF(ISO)4E arose specifically in angiosperms. It also incorporates a statistical framework to strengthen conclusions.

3.2.3 Eukaryotic Translation Initiation Factor Isoform 4E

Our collaborators, Patrick and Browning (2012), asked us to test their findings from 2012 that eIF(ISO)4E arose in angiosperms. Few plant sequences were available at the time and they proposed validating that earlier study as part of this collaboration. I identified the Pfam protein family PF01652 (Mistry et al., 2021) from *A. thaliana* sequence At5g35620. All sequences from the family were downloaded, processed, and visualised in a SSN (Figure 3.6A) to validate and filter by orthology.

I then applied Evolocuity (Hie et al., 2022) to determine mutational directionality. The analysis clearly shows evolutionary direction from eIF4E to the eIF(ISO)4E cluster (Figure 3.6B). With the origin of eIF(ISO)4E from eIF4E

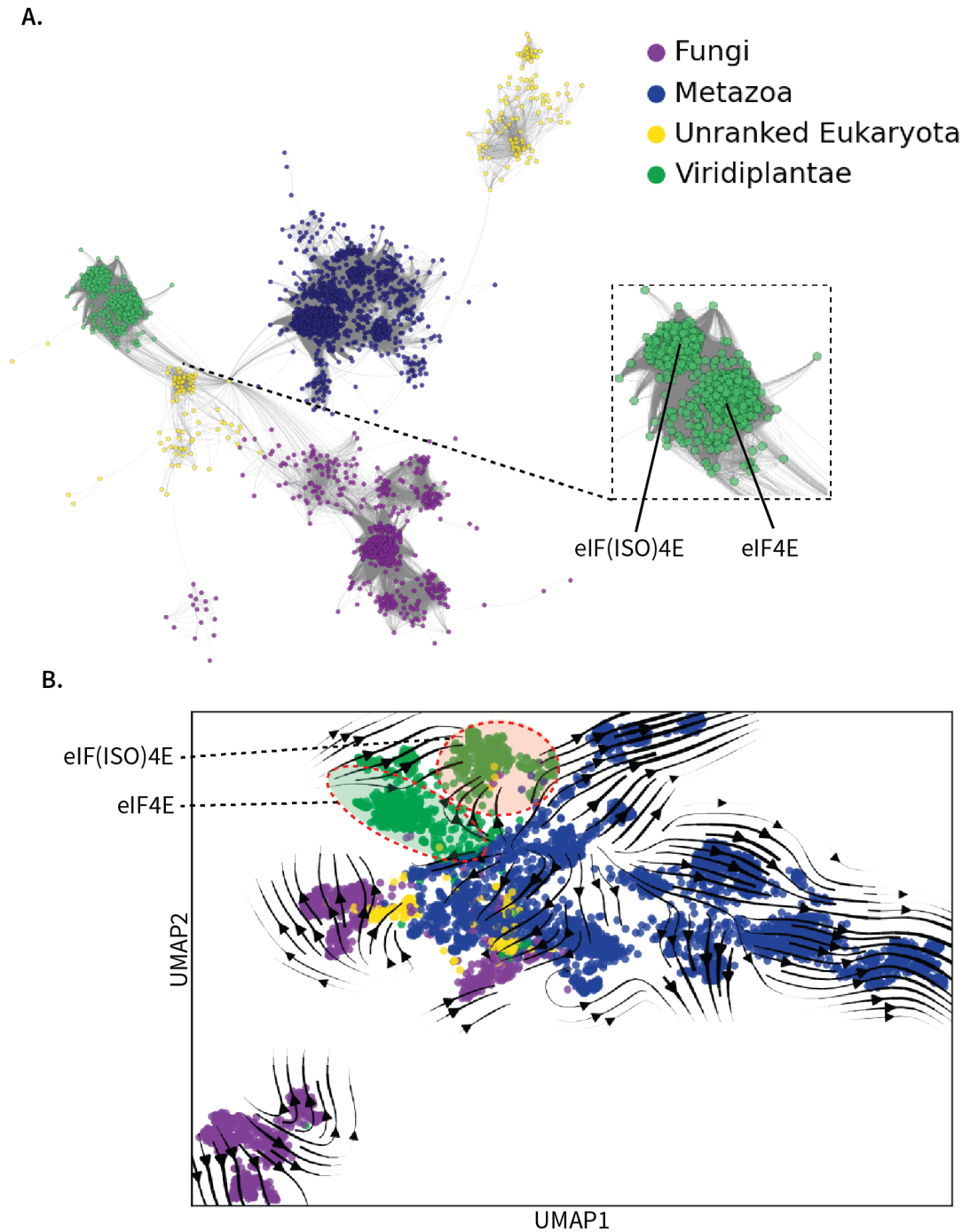


Figure 3.6: Network analysis of eIF4E and eIF(ISO)4E.

(A) The SSN coloured according to taxonomy with edges filtered to percentage identity greater than 52.5%, and nodes representing sequences. The cutout box shows the Viridiplantae only cluster independent of the network and identifies the eIF4E and eIF(ISO)4E subclusters. The SSN was clustered using the perfuse force directed OpenCL layout. (B) Evolocity network (KNN-based) coloured according to sequence kingdom, using UMAP. Sequences are represented as nodes in the network. Arrows show evolutionary velocity (direction and relative rate of change). The plant eIF4E and eIF(ISO)4E clusters are labelled.

established, I tested the statistical significance of eIF(ISO)4E presence or absence between angiosperms and non-angiosperms, and validated that the dataset size provided sufficient statistical power.

A one-tailed Mann-Whitney U-test examined whether non-angiosperms have significantly lower **eIF(ISO)4E** counts than angiosperms ($n = 43$ orders; 12 non-angiosperm, 31 angiosperm). The test yielded $U = 2.0$ with $p = 2.59 \times 10^{-7}$. A one-tailed Fisher's exact test for **eIF(ISO)4E** presence frequency also showed significant difference ($p = 2.09 \times 10^{-9}$). Power analysis using observed effect sizes confirmed adequate sample size for statistical power at $\alpha = 0.05$, with Cohen's $d = 0.9$ (estimated power = 1.00) and $\phi = 6.18$ (estimated power = 1.00). These tests strongly support that **eIF(ISO)4E** is absent from non-angiosperms. The single count in Mamiellales is an outlier that warrants further investigation to determine whether it represents annotation error, a pseudogene, or an independent evolutionary event.

Table 3.3 reveals that several angiosperm clades show substantially higher **eIF4E** counts than **eIF(ISO)4E** counts. To test whether this difference is statistically significant, I performed pairwise comparisons across all angiosperm orders using a two-tailed Wilcoxon signed-rank test for non-zero median difference between enzyme counts ($n = 31$ orders, $\alpha = 0.05$). The test statistic of 93.0 ($p = 3.64 \times 10^{-3}$) demonstrates that **eIF4E** counts are significantly higher than **eIF(ISO)4E** counts across the dataset.

I then tested whether any individual angiosperm orders showed significantly different **eIF(ISO)4E** retention compared to the overall pattern. Each order with ten or more sequences was compared against all remaining orders using two-tailed Fisher's exact tests on 2×2 contingency tables (Table 3.4). After correction for multiple comparisons using the Benjamini-Hochberg **False Discover Rate (FDR)** procedure at $\alpha = 0.05$, no individual orders showed statistically significant deviation from the expected **eIF(ISO)4E** proportion. Orders ranged from 21.4% retention in Asparagales to 56.2% in Malpighiales, but sample sizes within individual orders limited statistical power to detect these differences as significant.

These three case studies demonstrate structural and sequence-based approaches to protein classification, each revealing both capabilities and limitations. The Discussion synthesises technical and biological findings across all three studies and identifies the methodological gap that emerges from their collective limitations.

Table 3.3: Distribution of eIF4E and eIF(ISO)4E across plant orders

Order	eIF4E	eIF(ISO)4E	Total	Ratio
<i>Non-angiosperms</i>				
Chloropicales	1	0	1	—
Mamiellales	6	1	7	1:0.17
Chlamydomonadales	8	0	8	—
Sphaeropleales	2	0	2	—
Chlorodendrales	2	0	2	—
Chlorellales	3	0	3	—
Trebouxiales	1	0	1	—
Klebsormidiales	1	0	1	—
Charales	1	0	1	—
Marchantiales	2	0	2	—
Funariales	8	0	8	—
Selaginellales	3	0	3	—
<i>Angiosperms – High retention ($\geq 45\%$ eIF(ISO)4E)</i>				
Malpighiales	14	18	32	1:1.29
Fagales	6	7	13	1:1.17
Rosales	19	20	39	1:1.05
Alismatales	5	5	10	1:1.00
Malvales	40	38	78	1:0.95
Lamiales	14	13	27	1:0.93
Fabales	33	28	61	1:0.85
<i>Angiosperms – Moderate retention (35–45% eIF(ISO)4E)</i>				
Cucurbitales	10	8	18	1:0.80
Ericales	8	6	14	1:0.75
Sapindales	11	7	18	1:0.64
Caryophyllales	7	4	11	1:0.57
Brassicales	79	43	122	1:0.54
<i>Angiosperms – Low retention ($< 35\%$ eIF(ISO)4E)</i>				
Poales	106	57	163	1:0.54
Ranunculales	13	6	19	1:0.46
Solanales	66	28	94	1:0.42
Asterales	30	10	40	1:0.33
Zingiberales	18	6	24	1:0.33
Asparagales	11	3	14	1:0.27
<i>Angiosperms – Insufficient data ($n < 10$)</i>				
Cornales	2	5	7	1:2.50
Apiales	1	2	3	1:2.00
Proteales	2	3	5	1:1.50
Gentianales	2	3	5	1:1.50
Amborellales	1	1	2	1:1.00
Vitales	3	3	6	1:1.00
Celastrales	2	2	4	1:1.00
Oxalidales	1	1	2	1:1.00
Myrtales	5	4	9	1:0.80
Arecales	3	2	5	1:0.67
Laurales	2	1	3	1:0.50
Trochodendrales	4	2	6	1:0.50
Saxifragales	3	1	4	1:0.33

The ratio column expresses the number of eIF(ISO)4E sequences per eIF4E sequence. Orders with fewer than 10 total sequences were excluded from statistical analysis due to insufficient sample size.

Table 3.4: Statistical analysis of eIF(ISO)4E retention across angiosperm orders

Order	n	Observed	Deviation	p (adj)	Significant
Asparagales	14	21.4%	−17.9%	0.558	No
Zingiberales	24	25.0%	−14.3%	0.558	No
Asterales	40	25.0%	−14.3%	0.397	No
Solanales	94	29.8%	−9.5%	0.397	No
Ranunculales	19	31.6%	−7.7%	0.764	No
Poales	163	35.0%	−4.3%	0.558	No
Brassicales	122	35.2%	−4.1%	0.635	No
Caryophyllales	11	36.4%	−2.9%	1.000	No
Sapindales	18	38.9%	−0.4%	1.000	No
Ericales	14	42.9%	+3.6%	0.887	No
Cucurbitales	18	44.4%	+5.1%	0.764	No
Fabales	61	45.9%	+6.6%	0.558	No
Lamiales	27	48.1%	+8.8%	0.635	No
Malvales	78	48.7%	+9.4%	0.397	No
Alismatales	10	50.0%	+10.7%	0.728	No
Rosales	39	51.3%	+12.0%	0.474	No
Fagales	13	53.8%	+14.5%	0.635	No
Malpighiales	32	56.2%	+16.9%	0.397	No

Each order was compared against the remaining orders using Fisher’s exact test on a 2×2 contingency table, with a two-tailed alternative hypothesis. The null hypothesis assumed no difference in eIF(ISO)4E proportion between the focal order and all other orders. P-values were corrected for multiple comparisons using the Benjamini-Hochberg FDR procedure at $\alpha = 0.05$. Only orders with ≥ 10 total sequences were included. Deviation represents the difference between observed and expected proportions, with negative values indicating lower retention than expected.

3.3 Discussion

3.3.1 Technical findings

Structural modelling capabilities and limitations

The structural modelling approach successfully identified mechanistic explanations for reduced PSY activity without complete loss of function. AF2 and ESMFold produced concordant models with high confidence scores (pLDDT 87–88), enabling detailed analysis of mutation impacts despite the absence of experimental crystal structures for plant PSY. The ability to rapidly model multiple variants and compare their structures enabled mechanistic insights that would be difficult to obtain experimentally, making this approach well suited for similar applications, particularly for mechanistic hypothesis development.

Each mutation affects PSY through distinct structural mechanisms. The M266I substitution directly impacts the substrate binding pocket through steric effects. The A352T mutation disrupts protein fold stability through introduction of polarity into a hydrophobic region and potential helix-helix clashes. The P178S mutation affects catalytic residue positioning through increased backbone flexibility. Each mechanism explains why the mutants show reduced but not complete loss of function.

The high conservation of all mutation sites across 116 plant species validates their functional importance and suggests these positions represent critical structural constraints for PSY activity. Conservation also indicates these findings are likely applicable across PSY in Viridiplantae, not just *A. thaliana*, supporting *A. thaliana* as a suitable model organism for this study.

Despite the success of this structural modelling approach, important limitations remain. AF2 and ESMFold generate static structures and cannot sample the conformational landscape that may be critical for understanding PSY's catalytic cycle or how these mutations affect protein dynamics (del Alamo et al., 2022). Additionally, these models cannot capture electrostatic field changes that likely impact protein-protein binding interfaces. Electrostatic interactions play a central role in protein-protein recognition, particularly during the initial encounter phase where long-range electrostatic forces guide binding partners toward productive orientations (Kahler et al., 2020). Deleterious mutations at protein-protein interfaces tend to destabilise the electrostatic component of binding energy, even when they do not obviously disrupt net charge, hydrophobicity, or hydrogen bond networks (Teng et al., 2009). This suggests that subtle electrostatic perturbations can impair binding through mechanisms that

are not apparent from static structural analysis alone. The M266I and A352T variants reduce OR interaction despite showing no obvious structural impediment at the predicted binding interface. These observations are consistent with electrostatic changes or conformational dynamics that static models fail to capture.

Sequence-based evolutionary analysis capabilities and limitations

Constructing high-quality amino acid phylogenies with more than approximately 500 sequences with high replication becomes computationally prohibitive (based on testing). The *Evolocity* analysis provided an efficient alternative for detailed evolutionary analysis without the computational overhead of traditional phylogenetics. For *DXO1*, I identified evolutionary roots, demonstrated *DXO1* divergence from fungal and metazoan homologues, and determined where that divergence occurred by incorporating basic statistical charting with an established cladogram. This approach also provided a robust method for classifying the two plant forms of *DXO1* based on serine-rich N-terminal domain presence.

The *DXO1* study was the only case in which I used a purpose-built orthologue-finding tool, in this case Hmmscan (Potter et al., 2018). Hmmscan requires a subset of sequences to already be classified. As I was using it to identify the clan and attempt to locate the last common prokaryotic ancestor, there were already well-classified Metazoa, Fungi and Prokaryota sequences available. However, it was not able to classify plant *DXO1* into those with and without a serine-rich N-terminus. Other orthologue-identifying tools are available but each has drawbacks and situational dependencies. For example, tools such as OrthoFinder come with high computational overheads, bench-marked at a run time of 1.8 days for just 256 species (Emms and Kelly, 2019). Exploration of the dataset with these tools is also not always possible, unlike a SSN (Atkinson et al., 2009; Copp et al., 2018) which can be explored using network analysis tools such as Cytoscape (Shannon et al., 2003). Automated tools are not always best suited for the task and often do not provide the flexibility to explore data in the same way as manual approaches.

The combination of *Evolocity* network analysis with SSNs provided a computationally efficient approach for addressing specific evolutionary questions in the eIF4E/eIF(ISO)4E study. By classifying unlabelled sequences using the SSN and control sequences from *A. thaliana*, then validating through *Evolocity* analysis and statistical testing, I achieved robust classification without the computational burden of full phylogenetic reconstruction. The statistical frame-

work (Mann-Whitney U-tests, Fisher's exact tests, and power analysis) provided rigour that was absent from earlier studies with limited sampling.

While effective and efficient, these sequence-based approaches remain limited by their reliance on sequence similarity for classification. The *Evolocity* protein language model approach essentially creates a more sophisticated SSN, functioning as a pseudo-structural similarity network, but still relies on correlations at the sequence-structure level for classification. Additionally, dimensionality reduction distorts the network, which limits interpretation, particularly at fine scales, while the correlation between sequence identity and functional conservation is imperfect. Despite these constraints, the method provides an accessible and efficient means of answering targeted evolutionary questions about protein family evolution. To better understand the evolutionary relationships between sequences would require a full phylogeny.

In the *DXO1* study, the *Evolocity* pseudo-time analysis identified two evolutionary roots rather than one (Figure 3.5B) with one within Fungi and another within Viridiplantae. This pattern arises when either the true common ancestor is absent from the dataset, this shows as two pseudo-roots assigned to the most basal sequences within each disconnected clade (Hie et al., 2022), or where there were two independent origins. The absence of detectable prokaryotic homologues ($E\text{-value} \leq 1$) is consistent with a single ancient origin followed by divergence so extensive that connecting sequences are either extinct or not yet sampled. The Archaeplastida-Opisthokonta divergence occurred approximately 1.5–1.8 billion years ago (Parfrey et al., 2011; Eme et al., 2014), sufficient time for sequence similarity to erode below detection thresholds. Alternatively, sequences from under-sampled eukaryotic supergroups such as Excavata or Amoebozoa might bridge the gap if adequately represented (Yoon et al., 2008; Hampl et al., 2009). Independent origins through convergent evolution cannot be excluded but appear less parsimonious given the shared catalytic domain architecture across kingdoms. The dual pseudo-root pattern is therefore best interpreted as reflecting deep divergence from a common eukaryotic ancestor rather than independent evolutionary origins.

The SSN approach was sufficient for distinguishing eIF4E from eIF(ISO)4E because the identity threshold was relatively high ($>52.5\%$). More sophisticated techniques may be required to understand differences between enzymes that have not been experimentally classified at a finer scale, particularly when examining gene duplication events, that are common in plants, that may not map cleanly onto sequence similarity, nor show functional changes that support retention after duplication.

The methodological gap

These three case studies collectively reveal a methodological gap in protein classification. Structural modelling excels at explaining mutation mechanisms through steric, stability, and flexibility effects, but cannot capture electrostatic field changes that influence protein-protein recognition, while electrostatic charge interactions impacting structural folds can only be inferred. Sequence-based evolutionary analysis efficiently handles large datasets and identifies evolutionary directionality, but classification relies on sequence similarity thresholds that may not reflect functional differences.

This gap becomes apparent in cases where functional differences may arise from properties that neither static structures nor sequence similarity can capture. In the **PSY** case, the M266I and A352T mutations reduce **OR** interaction despite no obvious structural impediment, suggesting electrostatic or dynamic effects not captured by current methods. In the **DXO1** case, classification succeeded because the serine-rich N-terminal domain provides an unambiguous sequence marker, but this approach would fail for functional variants lacking such distinctive features. In the **eIF4E/eIF(ISO)4E** case, the clear sequence divergence (>52.5% identity threshold) made **SSN** classification tractable, but the chapter explicitly notes that finer-scale classification remains challenging.

The common thread is that current computational methods struggle when functional differences arise from electrostatic properties of protein active sites or binding interfaces rather than from sequence divergence or gross structural changes. This limitation motivates development of methods that can directly analyse and compare electrostatic fields across protein families.

3.3.2 Biological findings

Mutation mechanisms in phytoene synthase

The **PSY** mutations reduce carotenoid pathway flux through altered substrate binding (M266I), disrupted protein stability (A352T), and impaired catalytic geometry (P178S). Importantly, differential effects on protein-protein interactions were observed. The psy-4 (M266I) and psy-90 (A352T) variants showed reduced interaction with **OR**, while psy-90 uniquely showed severely impaired interaction with **GGPPS11**. The psy-130 (P178S) variant interacted normally with both **OR** and **GGPPS11**, suggesting its reduced activity stems purely from catalytic impairment.

The differential effects on protein interactions correlate with protein stability. Western blot analysis revealed that psy-4 and psy-90 had reduced **PSY** pro-

tein levels and corresponding reductions in OR protein levels (Bowerman et al., 2025). In contrast, psy-130 and psy-145 maintained normal PSY and OR protein levels. The parallel changes in both proteins suggest that PSY-OR interaction is crucial for mutual protein stability.

The mutations fine-tune carotenoid biosynthesis without complete pathway disruption. Specifically, they reduce accumulation of linear cis-carotenes including phytoene, phytofluene, ζ -carotene isomers, and neurosporene by 29–65%, while maintaining normal levels of downstream xanthophylls and β -carotene. Such selective reduction in pathway intermediates is sufficient to decrease production of the carotenoid-derived signal causing *ccr2* phenotypes.

Neo-functionalisation in decapping exoribonuclease and translation initiation factors

Both the DXO1 and eIF4E/eIF(ISO)4E studies reveal parallel patterns of neo-functionalisation following gene duplication or domain acquisition in angiosperms.

The evolutionary analysis demonstrates that angiosperms possess a unique DXO1 homologue characterised by a serine-rich N-terminal domain that emerged early in angiosperm evolution. This domain enables nuclear interaction with RNMT1, activating m7G cap methylation in a plant-specific partnership that parallels the mammalian RNMT/RAM system (Xiao et al., 2023; Gonatopoulos-Pournatzis et al., 2011). Loss of the domain causes ribosome loading defects and impaired translation-dependent mRNA surveillance, including elevated EJC footprints and suppressed co-translational RNA decay (Ganguly et al., 2025). The mechanistic link between RNMT1 activation and ribosome loading remains to be established, but proper cap methylation is required for recognition by cap-binding proteins that initiate translation (reviewed in Sonenberg and Hinnebusch (2009); Maquat et al. (2010)). This contrasts with fungal Rai1 and mammalian DXO, which directly degrade non-canonically capped RNAs through their exoribonuclease activity (Jiao et al., 2013; Xiang et al., 2009). The domain's prevalence across all streptophyte orders, combined with the observed clustering patterns, supports loss-of-function as the null hypothesis for streptophyte sequences lacking the serine-rich N-terminal domain.

The eIF4E/eIF(ISO)4E analysis provides strong support for eIF(ISO)4E arising in angiosperms as initially proposed by Patrick and Browning (2012), but with substantially higher confidence through expanded taxonomic sampling. The Evolocuity analysis demonstrates clear evolutionary directionality from eIF4E to eIF(ISO)4E, supporting an origin through gene duplication coinciding with

angiosperm emergence. Existing functional studies (Martínez-Silva et al., 2012) support neo-functionalisation following this duplication event. The statistical tests confirm eIF(ISO)4E emergence in angiosperms with very high significance ($p < 1 \times 10^{-6}$ for both tests).

The parallel pattern of angiosperm-specific innovations in regulatory control over translation and RNA metabolism suggests that the radiation of flowering plants was accompanied by elaboration of post-transcriptional regulatory mechanisms. Both DXO1 and eIF(ISO)4E provide additional regulatory flexibility compared to their ancestral or non-angiosperm counterparts.

Bimodal sequence length distribution in metazoan DXO1

Metazoan DXO1 homologues exhibit a bimodal length distribution (Figure 3.4A), in contrast to the unimodal distributions observed in fungi and plants. The approximately 200 aa difference between the two metazoan populations mirrors the length contribution of the serine-rich N-terminal domain in angiosperms, suggesting that some metazoan lineages may have acquired or retained additional regulatory domains.

This pattern could reflect several evolutionary scenarios. Mammalian DXO has been characterised as functionally distinct from fungal Rai1, possessing both decapping and exoribonuclease activities that fungal homologues lack (Jiao et al., 2013; Chang et al., 2012). The longer metazoan population may represent lineages where additional domain acquisition enabled functional diversification beyond the ancestral RNA surveillance role. Alternatively, the shorter population may represent secondary loss of regulatory domains in lineages where streamlined function proved advantageous.

The taxonomic distribution of these two length classes would clarify whether the bimodal pattern reflects a single divergence event early in metazoan evolution or multiple independent domain gain or loss events. If the longer sequences cluster within a monophyletic group, this would support a single acquisition event followed by retention in that lineage. Conversely, scattered distribution of longer sequences across metazoan phylogeny would suggest either multiple independent acquisitions or ancestral presence followed by differential loss. Mapping sequence length onto a metazoan phylogeny, combined with domain architecture analysis using tools such as InterProScan, would distinguish between these possibilities and identify which domains contribute to the length difference.

DXO1: Non-plant sequences with serine-rich N-terminal domains

Eight non-plant sequences possess serine-rich N-terminal domains comparable to those found in angiosperms (Table 3.2). These sequences span taxonomically disparate lineages with three fungal sequences from Basidiomycota and Chytridiomycota, and five metazoan sequences from Nematoda, Mollusca, and Platyhelminthes. Four hypotheses could explain this distribution.

First, these sequences may represent annotation artefacts. Computationally predicted gene models can incorrectly assign start codons or merge adjacent open reading frames, potentially generating spurious N-terminal extensions. However, the taxonomic distribution of affected sequences argues against systematic annotation errors affecting the same protein family across multiple independent genome projects.

Second, convergent evolution could have independently generated serine-rich N-terminal domains in multiple lineages. This explanation is unsatisfying for two reasons. No intermediate sequences exist showing progressive serine accumulation in N-terminal regions, which would be expected if convergent selective pressure drove domain evolution. Additionally, serine-rich domains serve specific functions in translational regulation and nuclear localisation (Xiao et al., 2023), making independent evolution of functionally equivalent domains in unrelated lineages improbable without shared selective pressure.

Third, the serine-rich domain may have been present in an ancient common ancestor and subsequently lost in most lineages. This scenario requires implausible numbers of independent loss events across fungi, metazoa, and non-angiosperm plants, making it the least parsimonious explanation.

Fourth, **Lateral Gene Transfer (LGT)** from plants could explain the presence of serine-rich domains in non-plant lineages. This hypothesis finds support in documented **LGT** patterns for several of the affected taxa. The fungal sequences derive from Entylomatales and Tilletiales, both orders within the smut fungi that are obligate plant pathogens. Plant-pathogenic fungi exhibit the highest rates of **LGT** among fungal lifestyles, with recent analyses claiming over 20,000 horizontally acquired genes across 829 fungal genomes (Liu et al., 2024). The intimate association between smut fungi and their plant hosts during infection provides opportunity for gene transfer. Similarly, plant-parasitic nematodes have acquired numerous genes from plants and plant-associated bacteria through **LGT**, particularly genes involved in plant cell wall degradation and carbohydrate metabolism (Danchin et al., 2010; Paganini et al., 2012). Whether the Rhabditida sequence in Table 3.2 are from species where any inter-

actions with plants could explain LGT could not be confirmed from available data.

However, the LGT hypothesis encounters difficulty with the remaining metazoan sequences. The cephalopod sequence derives from Octopoda, marine organisms with no plant association. While LGT has been documented in *Octopus bimaculoides*, the 12 identified transfer events involved marine bacteria rather than plants (Liu et al., 2018). The three Platyhelminthes sequences belong to Macrostomida, an order containing free-living marine flatworms such as *Macrostomum lignano* that feed on diatom algae rather than vascular plants (Wudarski et al., 2020). No clear mechanism exists for plant-to-flatworm gene transfer in these taxa.

The available data are insufficient to confidently distinguish between these hypotheses, and different explanations may apply to different sequences. The fungal sequences from obligate plant pathogens represent the most plausible LGT candidates, given established precedent for plant-to-fungus gene transfer in these intimate host-pathogen associations (Richards et al., 2009). By contrast, the cephalopod and flatworm sequences lack any obvious ecological connection to plants that would facilitate gene transfer. Convergent evolution, or contamination/data-basing issues represent more likely explanations for these taxa.

Robust identification of LGT events typically requires integration of multiple lines of evidence (Bock, 2010). These include phylogenetic incongruence between gene and species trees, atypical sequence composition, patchy taxonomic distribution, and ideally experimental validation to exclude contamination. The current analysis satisfies only the patchy distribution criterion. Compositional analysis comparing GC content and codon usage between putative transfers and their genomic backgrounds would provide additional discriminatory power. Phylogenetic analysis of the serine-rich domain itself, rather than the conserved DXO1 catalytic core, could determine whether non-plant sequences cluster with specific angiosperm lineages. Such clustering would support recent LGT, while basal positioning would suggest ancient divergence.

The small number of affected sequences (8 of over 2,000 analysed) limits the impact of this uncertainty on broader conclusions about DXO1 evolution. Regardless of their origin, these non-plant sequences with serine-rich N-terminal domains represent rare exceptions. They do not alter the fundamental observation that this domain arose predominantly within angiosperms and distinguishes plant DXO1 from its fungal and metazoan counterparts. The functional significance of serine-rich domains in non-plant contexts, should they prove genuine rather than artefactual, remains unexplored.

Differential retention of eIF(ISO)4E across angiosperm orders

The distribution of eIF4E and eIF(ISO)4E across angiosperm orders reveals consistent patterns that suggest biologically meaningful variation in paralogue retention. Across the 31 angiosperm orders examined, the overall proportion of eIF(ISO)4E was 39.3% (337 of 858 total sequences), with individual orders ranging from 21.4% (Asparagales) to 56.2% (Malpighiales). This variation indicates that selective pressures on eIF(ISO)4E maintenance differ across lineages.

No individual orders showed statistically significant deviation from the expected eIF(ISO)4E proportion after FDR correction (Table 3.4). This result likely reflects limited statistical power rather than an absence of biological signal. Even orders with substantial deviations from the expected ratio, such as Asparagales (21.4% observed versus 39.3% expected), did not reach significance due to small sample sizes within individual orders. Fisher's exact test provides exact p-values regardless of sample size, but its power to detect true differences decreases as sample size decreases. The FDR correction further reduces the likelihood of detecting effects when testing 18 orders simultaneously. However, the consistent directional patterns observed across orders suggest that the variation is not random. Orders cluster into distinct groups based on their retention patterns, with high-retention orders consistently maintaining eIF(ISO)4E proportions above 45% and low-retention orders consistently falling below 35%. This clustering would be unlikely if the variation were purely stochastic.

The pattern of retention and loss provides insight into the evolutionary dynamics following gene duplication. Multiple angiosperm lineages show significantly reduced eIF(ISO)4E counts relative to eIF4E, suggesting selective loss in certain clades indicative of genome contraction following duplication events (Panchy et al., 2016). Importantly, this loss is never complete within any order sampled, indicating either ongoing selective pressure for retention in some species or insufficient time for complete loss.

Ecological and life history correlates of eIF(ISO)4E retention

The pattern of eIF(ISO)4E retention correlates with growth form and life history strategy across angiosperm orders. Orders characterised by high eIF(ISO)4E retention are dominated by woody perennial species. Fagales comprises of almost exclusively deciduous and evergreen trees including oaks (*Quercus*), beeches (*Fagus*), birches (*Betula*), and walnuts (*Juglans*) (The Angiosperm Phylogeny Group et al., 2016). Rosales contains many economically important fruit trees such as apples (*Malus*), pears (*Pyrus*), and stone fruits (*Prunus*), along-

side shrubs like roses (*Rosa*) (The Angiosperm Phylogeny Group et al., 2016). Malpighiales includes poplars (*Populus*) and willows (*Salix*), and Malvales contains large tropical trees including baobabs (*Adansonia*) and kapok (*Ceiba*) (The Angiosperm Phylogeny Group et al., 2016). These orders share characteristics of long generation times, complex developmental programs involving seasonal growth cycles, vegetative dormancy periods, and require correspondingly complex stress regulation and adaptation strategies (Singh et al., 2017).

In contrast, orders with low eIF(ISO)4E retention are dominated by herbaceous plants with shorter generation times. Asterales, which shows the lowest retention among well-sampled orders (25.0%, n=40), contains the Asteraceae, one of the two largest angiosperm families with over 25,000 species (Mandel et al., 2019). Asteraceae is predominantly herbaceous with many annual and biennial species. Solanales (29.8% retention) includes nightshades, tomatoes, and tobacco, which are largely herbaceous annuals or short-lived perennials (The Angiosperm Phylogeny Group et al., 2016). Poales (35.0% retention), the largest order in this dataset, comprises grasses (Poaceae), sedges (Cyperaceae), and rushes (Juncaceae). These are predominantly herbaceous, with many annual or short-lived perennial species adapted to open, high-light environments (The Angiosperm Phylogeny Group et al., 2016). The low-retention orders also contain many species with weedy or invasive life history strategies characterised by rapid reproduction, high seed output, and broad environmental tolerance (Baker, 1965). This may suggest that eIF(ISO)4E is associated with more complex, longer term stress management strategies that are not retained in short-lived species.

Three of the six low-retention orders are monocots (Poales, Asparagales, and Zingiberales), suggesting that monocot lineages may experience reduced selective pressure for eIF(ISO)4E maintenance compared to eudicots. This observation warrants further investigation with phylogenetically controlled analyses.

eIF(ISO)4E life history trade-off hypothesis

The observed patterns suggest a trade-off between life history strategy and translational regulatory complexity. Long-lived woody perennials face selective pressures that may favour maintenance of both eIF4E and eIF(ISO)4E. These species must regulate translation across multiple developmental contexts including seasonal growth initiation, dormancy, flowering, and stress responses that recur over decades (Sajeev et al., 2019). The availability of two translation initiation factor isoforms may provide flexibility in how mRNAs are selected for

translation under different physiological conditions. Stress-responsive translation often involves selective translation of specific mRNA subsets (Branco-Price et al., 2008), a process that could benefit from isoform-specific regulation.

Conversely, herbaceous species with rapid life cycles may experience selection for streamlined regulatory networks. When generation time is short and reproductive success depends on rapid growth and seed production, the metabolic costs of maintaining duplicated regulatory machinery may outweigh the benefits of translational flexibility (Lynch and Marinov, 2015). Annual plants complete their entire life cycle within a single growing season, reducing the need for complex regulatory responses to seasonal variation, or complex stress regulation. Many short lived plants will simply set seed when subjected to sufficient environment stressors (Shavrukov et al., 2017). The consistent reduction of eIF(ISO)4E in lineages dominated by such species suggests that translational simplification may be adaptive in these contexts.

This hypothesis generates testable predictions. If the trade-off is real, eIF(ISO)4E expression should show tissue-specific or condition-specific patterns in species that retain it, reflecting functional specialisation between isoforms. Transitions from perennial to annual habit within lineages should also correlate with eIF(ISO)4E loss or pseudogenisation. Future work examining eIF(ISO)4E expression across tissues and developmental stages in species representing different life history strategies could test these predictions directly.

eIF(ISO)4E: Caveats and alternative explanations

Several factors could influence the observed patterns independently of life history. Sampling bias may affect orders with greater agricultural importance, as crop species typically receive more sequencing effort. Poales, Solanales, and Brassicales contain many major crops, and differential annotation quality across genome projects could affect paralogue detection rates. Polyploidy is also common in angiosperms and could affect paralogue retention through mechanisms unrelated to functional selection on translation initiation (Van de Peer et al., 2017).

The phylogenetic structure of the data presents another consideration. High-retention orders cluster in certain regions of the angiosperm phylogeny, and the patterns observed may reflect shared ancestry rather than independent adaptation to similar ecological conditions. Phylogenetically independent contrasts or sister-group comparisons would help distinguish between these possibilities (Felsenstein, 1985).

Despite these caveats, the correlation between growth form and eIF(ISO)4E

retention appears across multiple independent lineages. Woody perennials appear in multiple rosid lineages including Fagales, Rosales, Malpighiales, and Malvales, suggesting the correlation with eIF(ISO)4E retention may reflect genuine biological signal rather than phylogenetic artefacts alone.

3.3.3 Experimental validation across case studies

Phytoene synthase: structural predictions validated

Our structural predictions were validated through multiple experimental approaches in Hou et al. (2024). Callus assays with chemical inhibition of phytoene desaturase (PDS) showed reductions in phytoene accumulation for all psy mutants. PDS catalyses the desaturation of phytoene to phytofluene in carotenoid biosynthesis; its inhibition causes phytoene to accumulate, providing a measure of PSY activity. Immunoprecipitation followed by *in vitro* activity assays confirmed reduced enzymatic activity for purified PSY variants. In *Escherichia coli* (*E. coli*) expression systems, psy-4 and psy-90 showed significantly reduced phytoene production, while psy-130 showed marginally lower levels, consistent with structural predictions.

Protein interaction assays validated disrupted binding between PSY variants and partner proteins. The psy-90 variant showed the most severe defects, failing to interact with OR in yeast growth assays and showing dramatically reduced interaction with GGPPS11. The psy-4 variant showed intermediate effects, with reduced but not abolished OR interaction. The psy-130 variant maintained normal interactions despite reduced catalytic activity.

The rescue of *ccr2* phenotypes demonstrates the biological relevance of reduced PSY activity. All PSY variants restored normal plastid structures in etio-plasts, chlorophyll accumulation during light-induced greening, and green leaf emergence under short photoperiods. The restoration correlated with reduced PIF3/HY5 ratios where PIF3 protein levels decreased while HY5 levels increased to wild-type levels in *ccr2* PSY double mutants. This correlation indicates that moderate reductions in carotenoid pathway flux can normalise carotenoid-derived signalling without compromising essential carotenoid functions.

While our structural models successfully predicted catalytic impairments, they could not fully explain the differential effects on protein-protein interactions, nor verify the underlying disruptive mechanisms. The psy-4 variant shows reduced OR interaction despite the M266I mutation being distant from predicted protein surfaces, suggesting that conformational changes or electrostatic field alterations not captured in static models may govern such interactions. Electrostatics may also demonstrate the impact of the A352T mutation

by introducing a charged moiety into the hydrophobic core of the protein, beyond simple steric impacts. This could be coupled with molecular dynamics simulations which intrinsically models electrostatic interactions.

The PSY case study differs from the DXO1 and eIF4E analyses in that it addresses structure-function relationships within a single enzyme rather than classification across protein forms. However, it reveals a critical limitation relevant to classification methodology. Structural models successfully predicted gross catalytic impairments but could not explain differential effects on protein-protein interactions. Classifying PSY variants by their functional consequences, including distinguishing those that impair substrate binding from those that disrupt regulatory interactions, would require more than static structural models. The psy-4 variant illustrates this directly. Its reduced OR interaction cannot be explained by obvious structural impediments at predicted binding interfaces, suggesting that electrostatic field perturbations or conformational dynamics not captured in static models may determine functional class membership. This observation motivates classification approaches that can detect these differences, particularly for protein families where functional variants cannot be distinguished by sequence divergence or gross structural changes alone.

Decapping exoribonuclease: evolutionary predictions validated

Findings from my co-authors in Ganguly et al. (Ganguly et al., 2025) support my evolutionary analysis and classification of plant DXO1s based on N-terminal serine frequency and their divergence in early angiosperms. This evolutionary divergence represents an angiosperm-specific adaptation in RNA surveillance mechanisms. Unlike fungal Rai1 and mammalian DXO, which destroy non-canonically capped RNA, plant DXO1 potentially facilitates RNA recapping (Jiao et al., 2013; Xiang et al., 2009).

My evolutionary analysis predicted that *A. thaliana* DXO1 diverged from fungal and mammalian homologues by acquiring the serine-rich N-terminal domain in angiosperms (Figure 3.5D). Transgenic complementation experiments directly test this prediction.

The DXO1(E394A/D396A) variant shows that the ancestral catalytic activity, conserved across kingdoms, is not required for *A. thaliana* DXO1's physiological functions. This variant rescues all major phenotypes despite lacking deNADding activity (Ganguly et al., 2025, Figure 2C–D). Conversely, DXO1(Δ N194) fails to complement defects despite retaining full catalytic activity. This functional dissociation confirms our prediction that the angiosperm-specific domain represents the primary functional innovation.

Target transcripts segregate into three non-overlapping populations (Ganguly et al., 2025, Figure 3E), further supporting evolutionary neo-functionalisation through domain acquisition (Ganguly et al., 2025). The ancestral deNADding function persists for one transcript set, while the N-terminal domain independently acquired regulatory control over mRNA surveillance pathways. This addition of function aligns with our finding that the serine-rich domain emerged early in angiosperm divergence (Figure 3.5D) rather than replacing existing functional domains.

Ribosome loading defects in *dxo1* mutants ((Ganguly et al., 2025, Figure 4B–C) represent a plant-specific phenotype absent from characterised fungal and mammalian knockouts (Ganguly et al., 2025). This finding is consistent with our identification of this domain emerging at the angiosperm divergence point. Testing non-angiosperm DXO1s should further validate these results. Polysome profiling in non-flowering plant species would reveal whether ribosome loading defects are specific to angiosperms harbouring the serine-rich N-terminal domain. Complementation experiments expressing non-angiosperm homologues in *A. thaliana* *dxo1* mutants would test whether these ancestral proteins can rescue ribosome loading. Electrostatic comparisons of the active sites could shed light on catalytic differences and, crucially, aid in classification of the enzyme family particularly within Viridiplantae.

eIF4E/eIF(ISO)4E: future experimental validation

The eIF4E/eIF(ISO)4E study has not yet undergone experimental validation equivalent to the PSY and DXO1 case studies. Several experiments would strengthen the conclusions and test the life history trade-off hypothesis.

Polysome profiling across species with different retention patterns. The life history hypothesis predicts that species retaining both eIF4E and eIF(ISO)4E use these isoforms differentially across developmental stages or stress conditions. Polysome profiling separates ribosome-bound mRNAs by sucrose gradient centrifugation could identify whether specific mRNA subsets preferentially associate with eIF4E versus eIF(ISO)4E ribosomes. Comparing polysome profiles between a high-retention woody perennial (e.g., *Populus trichocarpa*) and a low-retention herbaceous annual (e.g., *A. thaliana*) under matched conditions would test whether retention correlates with isoform-specific translational regulation. Additionally, profiling across seasonal transitions in perennial species would reveal whether eIF(ISO)4E contributes to dormancy-related translational control.

Pseudogene identification in low-retention lineages. The trade-off hypothesis predicts that **eIF(ISO)4E** loss in herbaceous lineages reflects relaxed selection rather than absence of the gene. Genome-wide searches for **eIF(ISO)4E** pseudogenes in Asterales and Solanales species would distinguish between gene loss (pseudogenisation) and failure to duplicate (never present). The presence of pseudogenes would support secondary loss under relaxed selection, while their absence would suggest that **eIF(ISO)4E** arose specifically in lineages where it was retained.

Complementation experiments. Expressing **eIF(ISO)4E** from high-retention species in *A. thaliana* *eif(iso)4e* mutants would test functional conservation. If **eIF(ISO)4E** from woody perennials rescues mutant phenotypes more effectively than **eIF(ISO)4E** from low-retention herbaceous relatives, this would suggest functional divergence correlating with life history.

3.3.4 Implications for protein classification methodology

The three case studies presented in this chapter demonstrate the utility of computational approaches for protein classification while revealing characteristic limitations. These limitations point towards methodological opportunities.

Structural modelling enabled identification of distinct mutation mechanisms in **PSY** without requiring experimental crystal structures. Sequence-based evolutionary analysis using **Evolocity** and **SSN** approaches efficiently handled large datasets (>5,000 sequences) that would be computationally prohibitive for traditional phylogenetics and a large body of work regardless if inferring a phylogeny using representative samples. The methods also successfully classify **DXO1** forms by N-terminal domain presence and confirmed the angiosperm origin of **eIF(ISO)4E**. In each case, computational predictions were validated or remain testable through experimental approaches.

However, each method has characteristic limitations that together define a methodological gap. Static structural models cannot capture electrostatic field changes or conformational dynamics, which likely explain cases where mutations disrupt protein-protein interactions without obvious structural impediment (**PSY**). Sequence-based classification relies on similarity thresholds that may not reflect functional divergence. It may succeed when divergence is large (>52.5% identity for **eIF4E** / **eIF(ISO)4E**) but struggle at finer scales or when functional variants lack distinctive sequence features.

The common limitation is that neither approach directly addresses electrostatic properties of protein active sites and binding interfaces. Electrostatic in-

teractions drive molecular recognition, substrate binding, protein folding, and catalytic mechanisms, yet current classification methods treat electrostatics only indirectly through sequence conservation of charged residues or static structural positioning. When functional differences arise from electrostatic field distributions rather than from significant sequence divergence or gross structural changes, existing methods are insufficient.

This gap motivates development of methods that can directly compare electrostatic fields across protein families. Chapter 5 develops such an approach using Fourier decomposition of electrostatic potentials calculated from protein structures. By representing electrostatic fields as frequency components, this method enables quantitative comparison of field patterns across protein variants. Such an approach can detect differences that sequence similarity and static structures cannot reveal, addressing the methodological gap identified across these three case studies.

Chapter 4

Sequence-based classification of plant nucleotide phosphatases

4.1 Publishing contribution

Information in this chapter derives from two sources. The plant nucleotide phosphatase classification has not yet been published and is foundational work for this thesis. The wheat SAL analysis was conducted as part of my co-authorship on [Bowerman et al. \(2025\)](#), where I performed the phylogenetic analysis that established the evolutionary framework for understanding SAL gene diversity in wheat. My contributions to that study included identifying SAL homologues across monocots, constructing and analysing phylogenetic trees, interpreting evolutionary patterns of gene duplication, and connecting these findings to the functional specialisation observed in field trials.

4.1.1 Significance statement

- **Methodological Approach:** This chapter develops and validates a framework combining [SSN](#) analysis for initial protein family classification with phylogenetic inference to resolve functional relationships among nucleotide phosphatase paralogues. Computational subcellular localisation prediction, co-expression analysis, and gene ontology enrichment support functional classifications of uncharacterised protein paralogues. The methodology is validated through application to wheat SAL genes, where phylogenetically-predicted functional divergence between two homologous groups (*Triticum aestivum* L. SAL1 (TaSAL1) and *Triticum aestivum* L. SAL2 (TaSAL2)) was confirmed by field trials across 15 environments showing opposing effects on yield.

- **Key Evolutionary Discovery:** Analysis of the plant nucleotide phosphatase family revealed distinct evolutionary trajectories for three paralogue groups:
 - SAL shows strong purifying selection with no retained duplicates from major polyploidy events, indicating gene dosage sensitivity. In wheat, independent SAL duplications in Pooideae produced the functionally divergent groups **TaSAL1** which is associated with core metabolic processes and **TaSAL2** which is likely specialised for stress-responsive pathways.
 - AHL underwent duplication and retention following the γ -polyploidy event, with the model organism *A. thaliana* appearing to have lost AHL- β .
 - A putative chloroplast-localised nucleotide phosphatase (CNP) family was identified with predicted association to photosystem II function.
- **Biological Significance:** The phylogenetic framework reconciled conflicting literature on wheat SAL manipulation by revealing that previous studies inadvertently targeted functionally distinct homologues simultaneously. Field trials demonstrated that **TaSAL2** deletion improved yield by 4–8% while **TaSAL1** deletion reduced it, establishing that evolutionary classification is essential before expensive genetic manipulation in polyploid crops. The broader analysis demonstrates contrasting functional constraints between related phosphatase paralogues with AHL's maintaining duplicate copies enabling neo/sub-functionalisation, while SAL's dosage sensitivity shows purifying selection that controls gene copy number. The putative CNP classification suggests an unrecognised regulatory role in photosynthetic processes.

4.2 Introduction

Plants integrate environmental signals through retrograde signalling pathways that coordinate nuclear gene expression with organellar status. One such pathway involves **3'-phosphoadenosine 5'-phosphate (PAP)**, a metabolic byproduct of sulphur assimilation that accumulates in chloroplasts during drought and high light stress (Estavillo et al., 2011; Chan et al., 2016). PAP moves from chloroplasts to the nucleus where it inhibits specific exoribonucleases, altering RNA metabolism and triggering stress-responsive gene expression. The enzyme SAL1¹ (adenosine phosphatase, EC: 3.1.3.7) regulates this pathway by

¹SAL1 does not have an expanded name available in the literature.

dephosphorylating PAP to Adenosine MonoPhosphate (AMP), and its activity is itself redox-sensitive, creating a retrograde signalling mechanism that links chloroplast redox state to nuclear gene expression (Chan et al., 2016). This PAP-mediated retrograde signalling pathway represents an important mechanism for coordinating stress responses across cellular compartments, making the SAL enzyme family relevant to understanding both fundamental plant biology and stress tolerance in crops.

AtSAL1 from *A. thaliana*, identified through homology with the yeast salt tolerance gene SchAL2 (Murguía et al., 1995), remains the best-characterised member of this enzyme family. Biochemical and genetic studies have established that AtSAL1 shows dual chloroplast and mitochondrial targeting (Estavillo et al., 2011), PAP phosphatase activity (Quintero et al., 1996), and functions as a key regulator of drought and high light stress responses (Chan et al., 2016; Wilson et al., 2009). However, AtSAL1 belongs to a broader nucleotide phosphatase family whose diversity and evolution remain poorly understood. *A. thaliana* encodes three additional SAL paralogues (AtSAL2/3/4) that remain largely uncharacterised (Gil-Mascarell et al., 1999; Gašparič et al., 2013; Gy et al., 2007), as well as the related enzyme AHL, which has received minimal attention beyond basic catalytic measurements and a single mutant study (Gil-Mascarell et al., 1999; Shin et al., 2019). The functional relationships among these paralogues, and whether they represent redundant copies or functionally specialised enzymes, remain unresolved.

Current understanding of nucleotide phosphatase evolution is limited by incomplete phylogenetic sampling. Zhao et al. (2019) presented the most extensive plant nucleotide phosphatase phylogeny to date (n=85), but this analysis excludes closely related SAL paralogues and samples a limited taxonomic range. These constraints prevent accurate functional classification of family members and obscure the evolutionary history of gene duplications.

A central question is whether different nucleotide phosphatase paralogues show contrasting evolutionary responses to Whole Genome Duplication (WGD). WGD events create opportunities for neo-functionalisation and sub-functionalisation, but the initial doubling of all genes also imposes constraints. Because WGD preserves stoichiometric ratios of proteins within cellular networks and signalling cascades, genes encoding components of multi-protein complexes or systems experience strong selection for retention to maintain dosage balance (Clark and Donoghue, 2018; Qiao et al., 2019). This dosage-balance hypothesis explains the non-random patterns of gene retention following WGD, with network-associated genes showing preferential preservation and stronger purifying selection than genes arising from other duplication modes (Qiao et al.,

2019). Over evolutionary time, compensatory drift may release paralogues from these constraints as expression levels diverge, eventually permitting gene loss or neo-functionalisation (Clark and Donoghue, 2018). These alternative trajectories produce distinct phylogenetic signatures where dosage-sensitive genes show rapid loss of duplicates while genes permissive of functional divergence retain multiple copies across lineages.

The γ -polyploidy event provides a natural experiment for testing these predictions. This whole-genome triplication occurred approximately 130–160 million years ago at the base of core eudicots (Jiao et al., 2012), providing sufficient evolutionary time for retention patterns to stabilise. Consistent with the dosage-balance model, WGD-derived gene pairs show exponential decay across plant lineages, with most duplicates lost within the first few million years following polyploidisation (Qiao et al., 2019). Retention of duplicates from an event as ancient as γ -polyploidy therefore indicates selective pressure favouring copy maintenance, while absence of retained duplicates across broad taxonomic sampling suggests dosage constraints that selected against functional divergence. No systematic analysis has examined whether nucleotide phosphatase paralogues show these contrasting patterns, although duplicate copies of SAL are found in many species, or what these patterns would indicate about functional constraints within the family.

Subcellular localisation provides important clues for functional classification, yet localisation patterns across the nucleotide phosphatase family remain unresolved. While the absence of transit peptides often indicates cytosolic localisation (Armbruster et al., 2009; Armenteros et al., 2019), systematic computational prediction across family members has not been reported. Experimental evidence for AHL localisation illustrates the challenges where computational predictions, shown later, across over 100 Viridiplantae sequences suggest predominantly cytosolic localisation with some dual targeting to the nucleus, while Shin et al. (2019) demonstrated nuclear localisation using AHL-GFP transgenic plants in *A. thaliana*. However, transformation techniques can produce localisation artefacts (Millar et al., 2009), and single-species studies may not represent the broader family. This highlights the need for comprehensive computational localisation analysis across taxonomically diverse sequences to inform functional classification.

SSNs combined with phylogenetic analysis provide a rigorous framework for protein functional classification in the absence of comprehensive experimental data. SSNs cluster proteins based on sequence similarity, enabling identification of orthologous relationships and discovery of previously unrecognised protein groups (Atkinson et al., 2009; Copp et al., 2018; Leebens-Mack

et al., 2019). Phylogenetic methods complement this approach by revealing evolutionary patterns indicative of functional divergence, such as differential retention of duplicates following polyploidy events (Thornton, 2004). When integrated with co-expression analysis and gene ontology mapping, these computational approaches enable large-scale functional classification of protein families. While computational predictions carry inherent uncertainty, experimental characterisation is also subject to error, and multiple independent lines of computational evidence applied across broad taxonomic sampling can provide robust functional inferences.

How valuable computational classification frameworks are ultimately depends on whether they can predict biologically meaningful differences. This is particularly important in polyploid crops, whose genomes have undergone multiple rounds of duplication that can produce functionally divergent gene copies (Qiao et al., 2019; Clark and Donoghue, 2018; Panchy et al., 2016). Bread wheat (*Triticum aestivum* (wheat) (L.) (*T. aestivum*)) exemplifies this complexity where its hexaploid genome arose through two hybridisation events, and individual gene families can contain multiple homoeologous copies with potentially distinct functions. Previous studies attempting to manipulate SAL in wheat produced conflicting results. Virus-induced gene silencing improved drought tolerance markers (Manmathan et al., 2013), and CRISPR-based knock-outs reported enhanced stress resistance based on osmotic shock experiments (Abdallah et al., 2025), while other CRISPR studies observed yield penalties under competitive glasshouse conditions (Mohr et al., 2022). These contradictory outcomes share a common limitation where no-one considered whether different SAL homologues might have distinct functions, and none measured impacts under field conditions where yield is the relevant metric for crop improvement.

This lack of evolutionary context creates interpretive problems that extend beyond individual studies. When genetic manipulation targets multiple gene copies simultaneously without understanding their functional relationships, unpredictable phenotypes can result from disrupting genes with opposing effects. While 5–7 SAL genes had been identified in different wheat varieties (Mohr et al., 2022), the evolutionary relationships among these copies remained unresolved. Phylogenetic classification is therefore essential before genetic manipulation to determine whether gene copies represent functionally redundant duplicates or specialised enzymes whose individual manipulation would produce distinct outcomes.

This chapter addresses these gaps through two complementary approaches. The first part develops a comprehensive functional classification of plant nucle-

otide phosphatases by constructing a taxonomically broad phylogeny spanning over 500 sequences from across Viridiplantae. This analysis integrates SSN clustering with phylogenetic inference, computational localisation prediction, co-expression analysis, and statistical comparison of sequence properties between paralogous clades. The aims are to:

- identify and classify previously unrecognised nucleotide phosphatase family members,
- determine patterns of gene duplication and retention that indicate functional specialisation, and
- predict functional divergence between paralogue groups through statistical analysis of sequence properties and evolutionary patterns.

The second part applies this classification framework to wheat SAL genes and tests whether Phylogenetically-predicted functional divergence corresponds to measurable phenotypic differences under field conditions. Field trials across 15 environments assessed whether different SAL homologues produce distinct yield effects when deleted, providing direct validation of whether sequence-based classification methods can reliably predict functional outcomes relevant to crop improvement.

This chapter focuses on sequence-based classification methods and their validation through phenotypic outcomes. The analysis examines whether statistical comparison of sequence properties such as length, aromaticity, and isoelectric point can support phylogenic determination of retained gene-duplication (and therefore functional divergence) giving rise to paralogous clades. Cases where these sequence-level metrics do not support gene-duplication in paralogous clade are outlined, but alternative classification approaches such as electrostatic field analysis are addressed in subsequent chapters.

4.3 Results

4.3.1 Classification of the plant nucleotide phosphatase family

A SSN was built to identify orthology and classify protein families through sequence similarity relationships (Atkinson et al., 2009; Copp et al., 2018; Leebens-Mack et al., 2019). This dataset incidentally contained sequences with sequence identity <15%. The SSN (Figure 4.1) successfully classified predicted SAL and AHL clusters matching the *A. thaliana* and Phytozome-derived (Goodstein et al., 2012) control sequences. A third previously uncharacterised cluster sharing homology with SAL and AHL was identified. This cluster isolated with the *A. thaliana* gene locus identifier At4G05090 which appears only once in the

literature and was not discussed (Kim et al., 2010).

This novel protein cluster was putatively named as **CNP** based on three functional prediction criteria:

1. Subcellular localisation prediction using DeepLoc2 and MitoFates (Thummuluri et al., 2022; Fukasawa et al., 2015) showed the majority (~ 69.5%) of sequences have predicted chloroplast localisation.
2. Co-expression and gene ontology analysis revealed co-expression with enzymes localised to the chloroplastic thylakoid lumen and associated with photosystem II (Table 4.4).
3. Tissue expression analysis showed primary expression in photosynthetic tissues in the cotyledons and developing leaves (Figure 4.7).

The **SSN** enabled functional classification through sequence similarity relationships between the different *A. thaliana* SAL homologues while filtering sequences unlikely to be orthologous. All six homologues – At**CNP**, At**AHL**, and AtSAL1/2/3/4 – were used as seed sequences for cluster classification. Notably, three *A. thaliana* sequences were classified as non-homologous to AtSAL1 at the scale investigated (sequence similarity <35%) (Figure 4.1).

Following **MSA** alignment optimisation, a careful alignment and the subsequent removal of transit peptides, a high-quality phylogeny was inferred for evolutionary and functional classification. To do this, an initial phylogeny was created and the closest branch below the SAL clade with >95 ultrafast bootstrap support was selected as the root for functional group delineation. The dataset was optimised to 500 representative sequences (plus the outgroup), using CD-HIT (Huang et al., 2010), for robust functional classification while maintaining computational feasibility. ModelFinder (Kalyaanamoorthy et al., 2017) was used to identify the highest performing evolutionary model – EXO_EHO+R9 (Bayesian Information Criterion (BIC) score 285583.624), while the second and third best performing models were JI+T+F+R9 (BIC score 287539.190) and LG+R10 (BIC score 287550.402) respectively.

The phylogeny was inferred with 400 replicates using the same dataset, all of which passed the Approximately-Unbiased statistical test (AU-test) to check for validity (Shimodaira, 2002). A consensus cladogram for the 20 highest scoring trees by log likelihood (lnL) + AU (Figure 4.2) was found to qualitatively best match the 1,000 Plant Transcriptome Initiative phylogeny topology, and the tree that most closely matched that consensus cladogram was picked as the final phylogeny.

The high-quality alignment and statistically well supported phylogeny derived from a massive pool of replicate trees provides high confidence in the final phylogeny and an excellent base for analysis. The final phylogeny (Figure

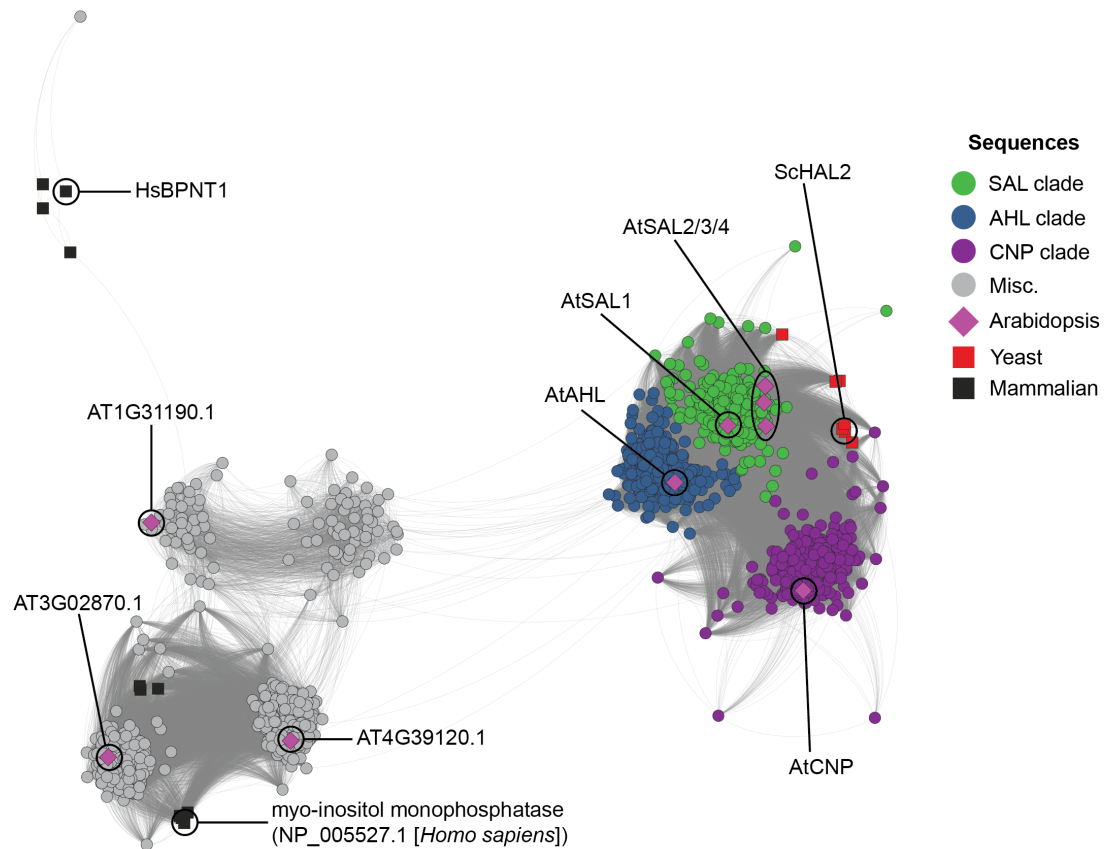


Figure 4.1: Sequence Similarity Network of SAL-homologue BLAST results. Edges (grey lines) are filtered to >85 alignment score, which was found to result in optimal clustering, and clustered by OpenCL Prefuse Force Directed Layout. Relative distances between groups of sequences indicate similarity, while the edges show relationships between sequences where the alignment score >85, the more edges the more related the groups of sequences. Phytozome (curated) sequences as diamonds provide reference for the One Thousand Plant Transcriptome Initiative (OneKP) sequences as circles, while also populating the dataset. *A. thaliana* sequences (pink) are controls for the clusters. The SAL cluster is green, AHL-related cluster is blue, and purple represents an unidentified putative nucleoside phosphatase cluster (CNP). The grey clusters are possible other nucleotide phosphatases not explored in this study and were repositioned for the purposes of the figure, as such their distance from the other clades is not indicative of their similarity.

4.3) reveals functional divergence patterns with CNP diverging first from the root, followed by AHL, and SAL last. The phylogeny places the origin of SAL in Streptophyta algae, AHL in Chlorophyta algae, and CNP in a common ancestor with Chromista and Viridiplantae.

Putative gene-duplication was identified in AHL and CNP identified as out-paralogous clades (Figure 4.3). The AHL clades showed statistically significant functional differentiation with very high differences in sequence length, aromaticity, and isoelectric point with moderate-to-large effect sizes (Table 4.1), supporting functional or subcellular environmental specialisation. Figure 4.4 shows the AHL paralogue only. The clade annotations are my own and will be used throughout the rest of this thesis. Clades denoted by major branch events

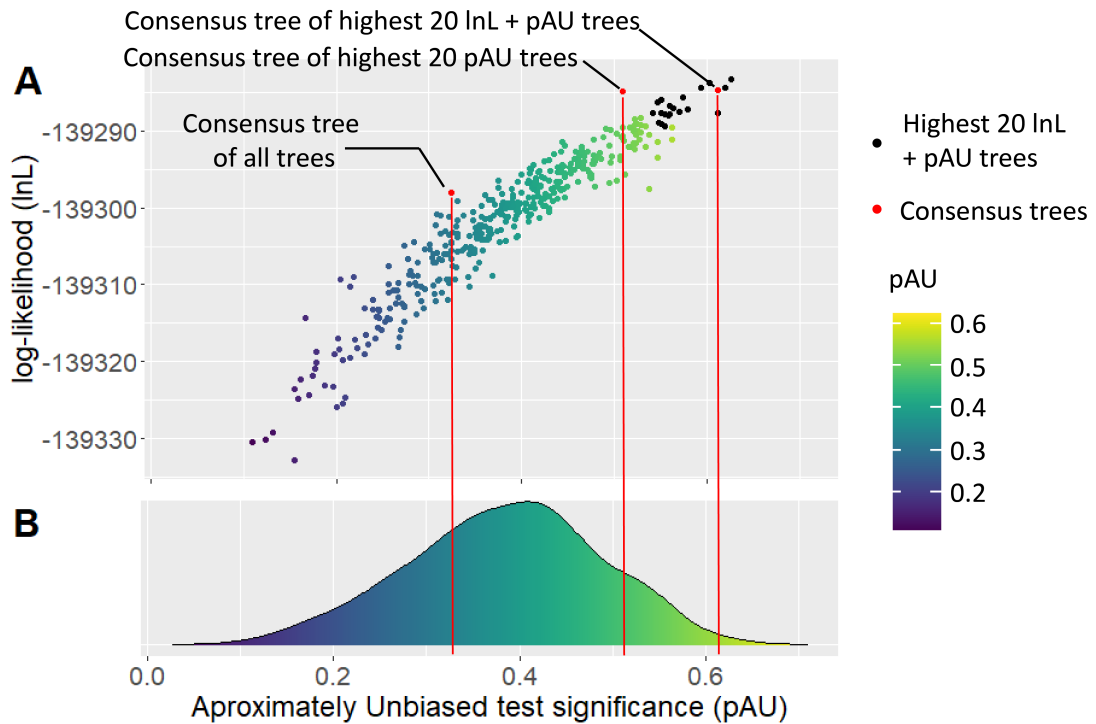


Figure 4.2: Quality score distribution of 400 replicate phylogenies. X-axis shows the Approximately Unbiased test significance (pAU) score. Y-axis shows log likelihood (lnL). Each dot represents the scores of a single phylogeny replicate coloured by pAU score with the following exceptions. The top 20 phylogenies by lnL and pAU are coloured black. Consensus trees are coloured red and labelled. Distribution of pAU scores is shown below the dot-plot, coloured by pAU. The Approximately Unbiased test defines any score >0 as a pass.

in the phylogeny as Alagal AHL (this clade contains chlorophyte algae specific sequences), Basal AHL (non-Eudicot Streptophyte sequences) and the outparalogous AHL clades AHL- α and AHL- β . Representative structures are shown for *Gossypium raimondii* (cotton originating in Peru) with GrAHL (AHL- α) and GrAHL2 (AHL- β), along side AtAHL. Clear structural differences between the AHL- α and AHL- β are visible.

The CNP outparalogous clade showed no significant functional differentiation (Table 4.2) with branch support <95 (Figure 4.3). A putative domain swap (alternatively a region of mutational rate-heterogeneity) was identified specific to the main CNP clade (Figure 4.5). The domain swap arises in or before Streptophyta algae but is absent in Glaucophyta, Chromista, or Rhodophyta CNP sequences. DeepLoc2 predicted plastid localisation for 69.2% of CNP sequences.

The classification framework successfully identified three distinct nucleotide phosphatase families and detected functional divergence between AHL outparalogous clades through statistical comparison of sequence properties. However, computational predictions require validation against phenotypic outcomes. Polyploid wheat provides an opportunity to test whether phylogenetic classi-

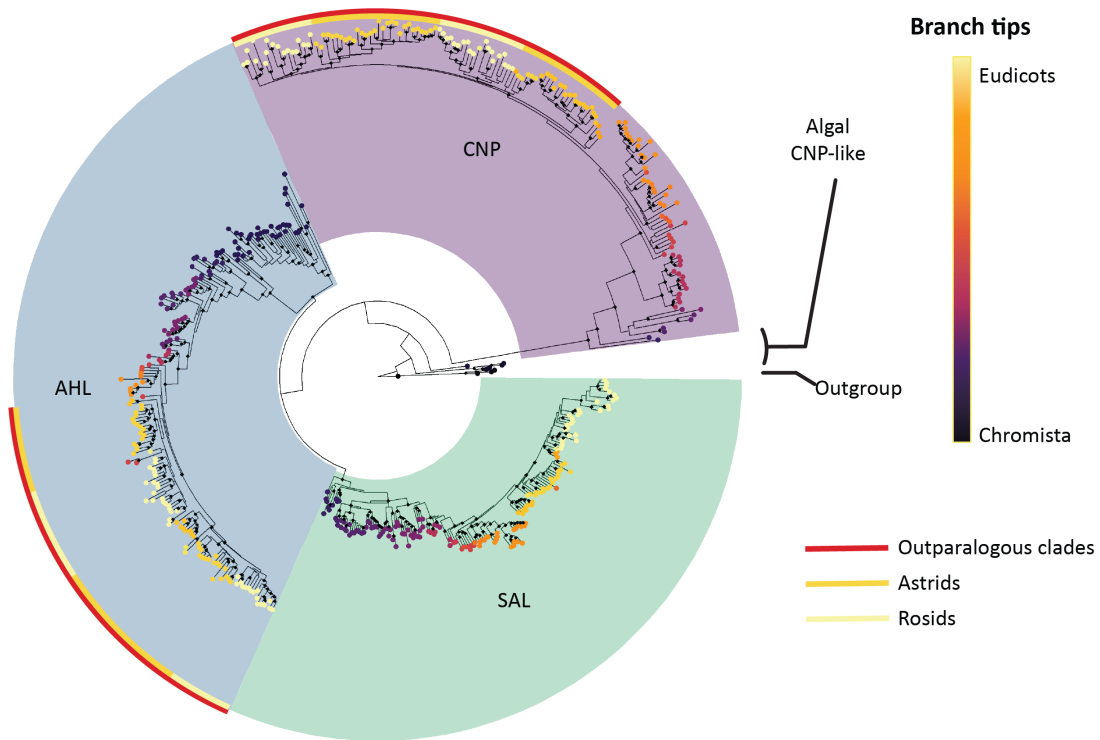


Figure 4.3: Plant nucleotide phosphatase circular layout phylogeny showing SAL, AHL and CNP clades. Paraloues are coloured by: AHL, blue; SAL, green; CNP, purple. Black dots at nodes indicate branches with >95 % confidence. The ultrafast bootstrap values for the AHL-SAL branch fell below the 95 % threshold and is not statistically considered a high-confidence branch. Nevertheless, this branch remains robust across the best-supported replicate trees, suggesting it is the most likely overall topology. The algal CNP-like sequences (*Chromista*, *Rhodophyta* and *Glaucophyta* sequences) sit adjacent to the outgroup. The mosses, hornworts and ferns clades in AHL do not match the species tree, likely an artefact of insufficient phylogenetic signal (i.e. the evolutionary relationships cannot be resolved given the sequence lengths relative to sequence variance for that few sequences). Branch length indicates relative evolutionary distance (unit-less). Putative outparalogous clades are annotated by red arcs with *Atrids* and *Rosids* shown with orange and yellow arcs respectively.

Table 4.1: Welch *t*-test using Cohen's *d* comparing the outparalogous AHL clades, n = 84*

Variable	Degrees of freedom and statistic	p-value	Significance	Effect size
Sequence length	t(43.66) = -3.69	0.00061	***	moderate
Aromaticity	t(73.19) = -5.46	<0.0001	****	large
Isoelectric point	t(66.31) = -5.17	<0.0001	****	large

*Significance threshold p = 0.05; ns = not significant, * = p 0.01, ** = p 1x10⁻³, *** = p 1x10⁻⁴, **** = p <1x10⁻⁴

fication can predict functional differences that manifest as measurable yield effects under field conditions.

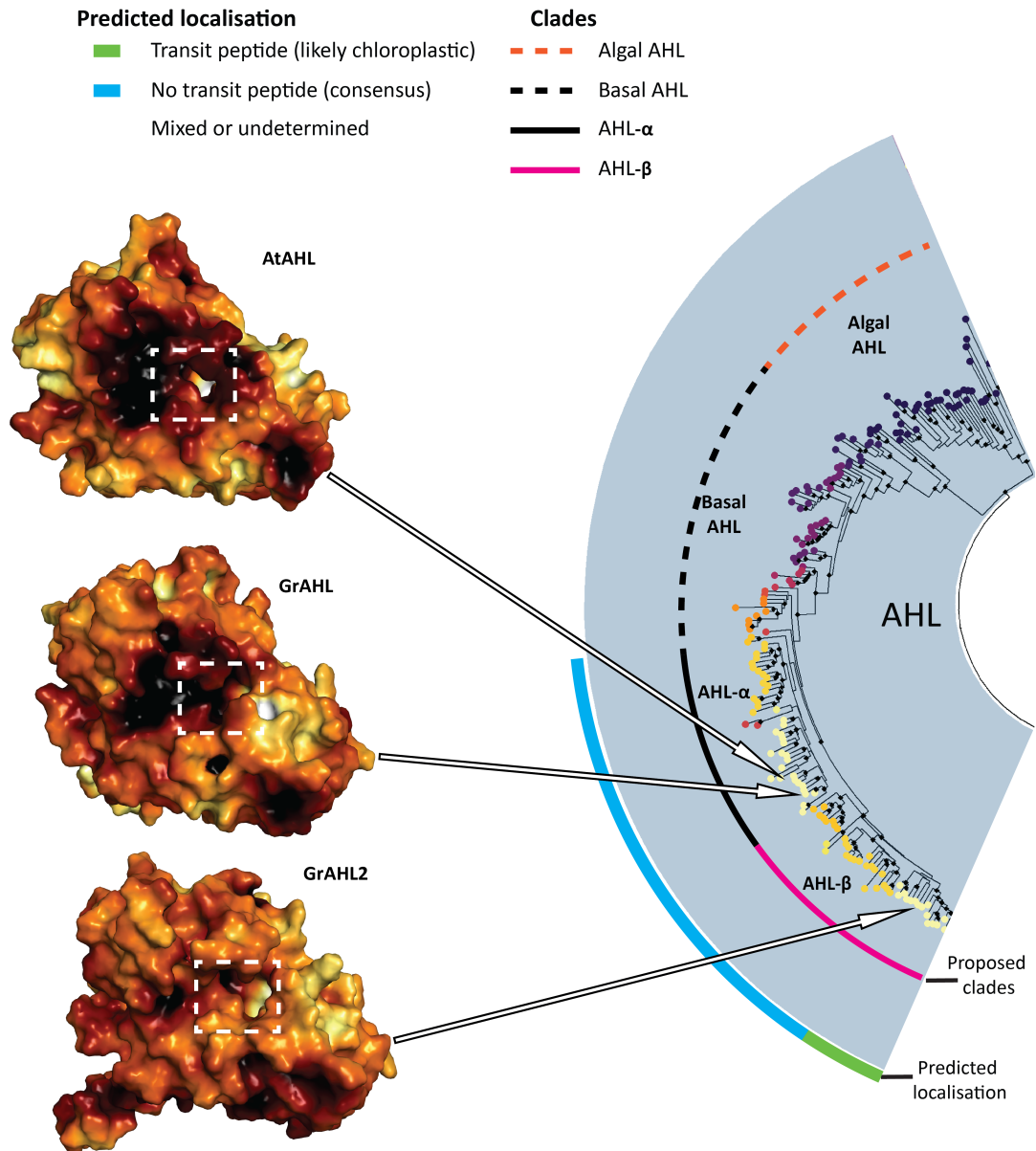


Figure 4.4: Surface charge representation of *Arabidopsis thaliana* and two *Gossypium raimondii* AHL models from the AHL- α and AHL- β clades respectively, at pH 7.3, related to the AHL phylogeny with localisation information. Blue arc indicates consensus of no transit peptide, green arc indicates transit peptides are present and were manually assessed as likely chloroplast targeting. Orange dashed arc indicates Algal AHL clade, black dashed arc shows Basal AHL clade, black solid arc shows AHL- α clade, and pink arc shows AHL- β clade. *Gossypium raimondii* AHL (GrAHL1) and GrAHL2 surface charge models with the putative active sites boxed in white. The structures are related back to where the sequences occur in the phylogeny. *A. thaliana* AHL (AtAHL) has been provided as a reference. The phylogeny has predicted localisation data from DeepLoc2 for the AHL- α and AHL- β paralogues. *G. raimondii* was chosen because of the species' established scientific interest, and it has sequences available in both clades. Surface charge is a unit-less value derived from the Boltzmann constant (k) multiplied by the temperature in degrees Kelvin divided by the amino acid charge at a set pH. Charge was calculated using default values at pH 7.9 using PBD2PQR-APBS.

4.3.2 Phylogenetic Classification of Wheat SAL Homologues

Phylogenetic analysis of monocot SAL sequences revealed that wheat (cv. Chara) possesses seven SAL gene copies organised into two distinct homologous groups

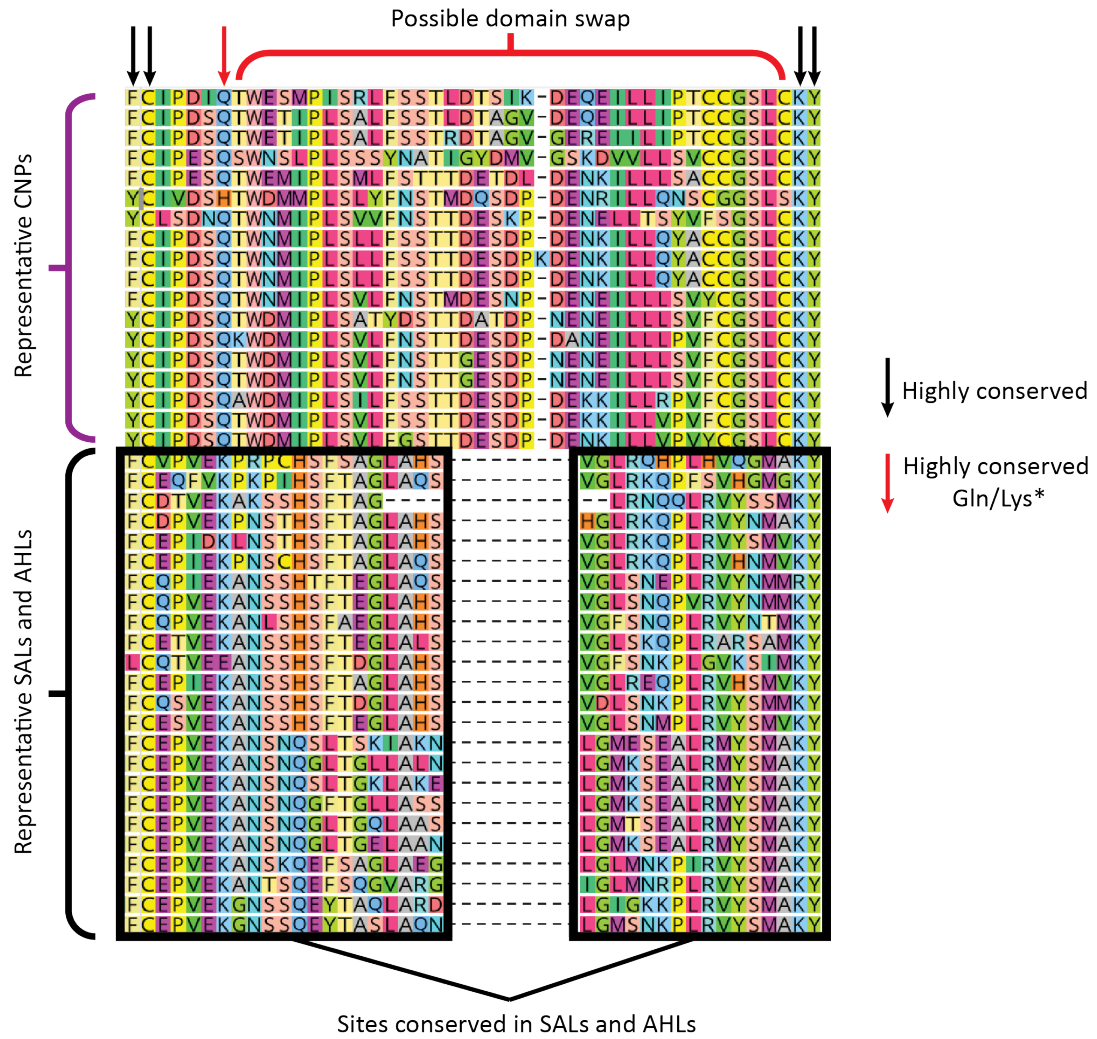


Figure 4.5: Local alignment of representative AHL, CNP and SAL sequences at location of CNP putative domain-swap or insertion. Sequences were aligned using MAFFT with default values, visualised in Geneious. Residues are using the standard single letter naming convention, and coloured by Cytoscape's default scheme. Dashes ('-') represent gaps. The alignment spans AtSAL1 residues F227 to Y263. Highly conserved residues are annotated with black arrows and red for a probable conserved residue: lysine in SAL and AHL and glutamate in CNP, noting only a single nucleotide mutation is required to mutate the residue between these two amino acids. Conservation of a site was determined off the full 500 sequence alignment. BLAST searches were conducted to try to identify the domain from other plant proteins, however likely due to its short length and age (likely resulting in sequence drift), no results were returned with e-values <0.5 (a 0.5 e-value is very poor and indicates that there is a one in two chance the sequence occurred in the database by chance).

Table 4.2: Welch *t*-test using Cohen's *d* comparing the outparalogous CNP clades, $n = 73^*$

Variable	Degrees of freedom and statistic	p-value	Significance	Effect size
Sequence length	$t(50.13) = -0.18$	0.86	ns	negligible
Aromaticity	$t(43.26) = -0.90$	0.37	ns	negligible
Isoelectric point	$t(52.00) = -0.33$	0.74	ns	negligible

*Significance threshold $p = 0.05$; ns = not significant, * = $p 0.01$, ** = $p 1 \times 10^{-3}$, *** = $p 1 \times 10^{-4}$, **** = $p < 1 \times 10^{-4}$

(Figure 4.6A). Bootstrap support values exceeding 95% confirmed robust separation between groups. These were designated as **TaSAL1**, located on chromosomes 5A/B/D, and **TaSAL2**, found on chromosomes 4A/7A/7D, with the 4A copy resulting from a known 4AL/7BS translocation (Devos et al., 1995).

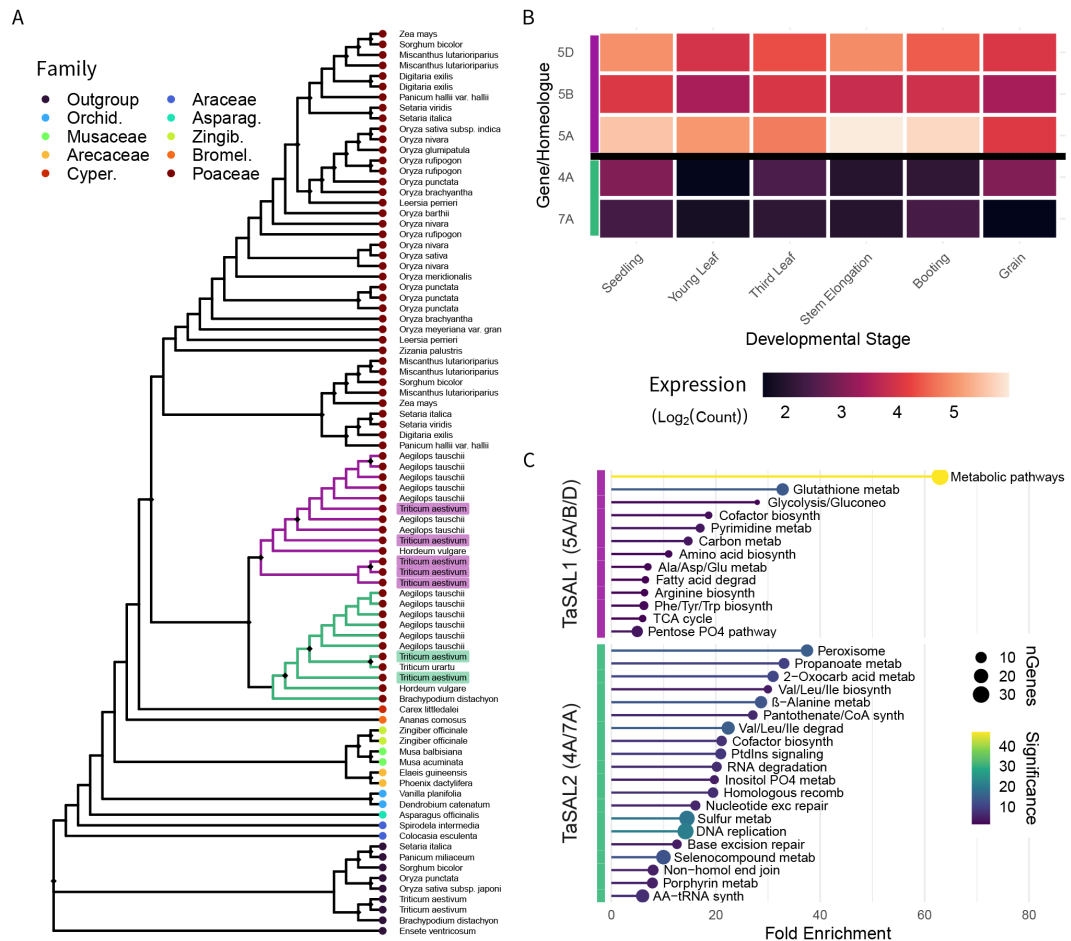


Figure 4.6: Wheat SAL gene diversity.

(A) Phylogeny of Liliopsida SAL sequences. Black diamonds indicate bootstrap support > 95% based on 1000 ultrafast bootstrap replicates. Purple, TaSAL1 (chr 5); green, TaSAL2 (chr 4A/7A/7D). (B) Expression of wheat SAL genes. (C) GO enrichment for TaSAL groups. Significance = $-\log_{10}(P\text{-value})$. Reproduced from (Bowerman et al., 2025), please see paper for full methods.

The phylogeny demonstrated that this duplication pattern arose from independent duplication events in both the Pooideae and Panicoideae subfamilies of Poaceae (Figure 4.6A). The conservation of both paralogous groups across grass evolution, rather than loss of one copy, indicated potential functional divergence rather than redundancy. This same duplication event gave rise to paralogous copies in *Hordeum vulgare* (barely) and *Aegilops tauschii* (Tausch's goat-grass). While *Aegilops tauschii* has multiple copies of both paralogues, *Hordeum vulgare* only has one copy. The Panicoideae duplications do not appear to be analogous. The duplications in Panicoideae occurred later and the

topography of the phylogeny in the family with poor bootstrap support from most branches suggests we can make little from this section of the phylogeny. This is likely due to over representation of genomes in the *Oryza* (rice) genus.

4.3.3 Expression and Co-expression Analysis Support Functional Divergence

RNA-seq analysis revealed distinct expression patterns between the phylogenetically-defined groups. *TaSAL1* homologues showed 10-fold higher expression than *TaSAL2* across developmental stages (Figure 4.6B). *TaSAL1* expression peaked in vegetative tissues during active growth phases, while *TaSAL2* showed lower but more variable expression across developmental stages.

Co-expression network analysis conducted by my co-authors revealed distinct functional associations for each phylogenetic group (Figure 4.6C). *TaSAL1*-associated networks showed enrichment for fundamental cellular processes including sulphur metabolism (11-fold enrichment), DNA repair (15-fold), and aminoacyl-tRNA pathways (62-fold). In contrast, *TaSAL2* networks enriched for metabolic flexibility pathways including carbon metabolism (14-fold) and amino acid metabolism (37-fold) (Bowerman et al., 2025).

4.3.4 Field Performance of Targeted SAL Deletions

To test whether the phylogenetically-predicted functional divergence manifested as phenotypic differences, field trials were conducted across 15 environments with targeted deletions of specific homologous groups. Lines with *TaSAL2* deletions (SAL2, SAL2.3) showed 4-8% yield improvements and enhanced water productivity compared to the Chara control. Conversely, *TaSAL1* deletions (SAL1.1, SAL1.2) consistently reduced yields. Double mutants targeting both groups (SAL1.1 × SAL2.1) showed severe yield penalties of 18% (Table 4.3).

Table 4.3: Yield, biomass, and water productivity.

Genotype	Yield (t ha ⁻¹)	Biomass (t ha ⁻¹)	Water productivity (kg ha ⁻¹ mm ⁻¹ water)
Chara	4.70 ± 0.03	9.41 ± 0.11	14.20 ± 0.10
SAL2	4.89 ± 0.03**	9.38 ± 0.11	14.34 ± 0.10***
SAL2.3	5.09 ± 0.20†	9.83 ± 0.46	14.29 ± 0.51†
SAL1.1	4.46 ± 0.03***	9.69 ± 0.11	13.28 ± 0.10***
SAL1.2	4.57 ± 0.03***	9.78 ± 0.11*	13.10 ± 0.10***
SAL1.1 × SAL2.1	3.85 ± 0.03***	8.02 ± 0.14***	11.68 ± 0.10***

Mean values of random effects ± SE. *, $P < 0.05$; *** $p < 0.001$ vs Chara. †Not significant in multienvironment trial model, but significant in trials present. This table has been reproduced from Bowerman et al. Bowerman et al. (2025).

The opposing yield effects of *TaSAL1* and *TaSAL2* deletions confirm that

phylogenetic classification identified functionally distinct gene copies. These results validate the sequence-based classification approach showing that the method succeeds when evolutionary divergence produces detectable differences in sequence properties, but cannot directly assess whether sequence differences translate to catalytic divergence.

4.3.5 *Arabidopsis thaliana* nucleotide phosphatase expression and Gene Ontology

AHL, CNP and SAL *A. thaliana* tissue expression from Klepikova et al. (2016) are shown in Figure 4.7. Expression of AtAHL is particularly high in the guard cells, while CNP is most highly expressed in mesophyll tissue in developing leaves. Gene ontology (Table 4.4 shows similar associations in SAL and AHL, while CNP is closely associated with the chloroplast and photosystem II. Expression (mRNA) of the three paralogues under drought conditions (GENEVESTIGATOR, experiment ID AT-00626) is presented here in Figure 4.8. SAL has a response as expected in its role as a drought sensor (Chan et al., 2016). Interestingly AtAHL has inverse expression when compared to AtSAL1 under the same conditions while AtCNP appears has no significant changes in expression.

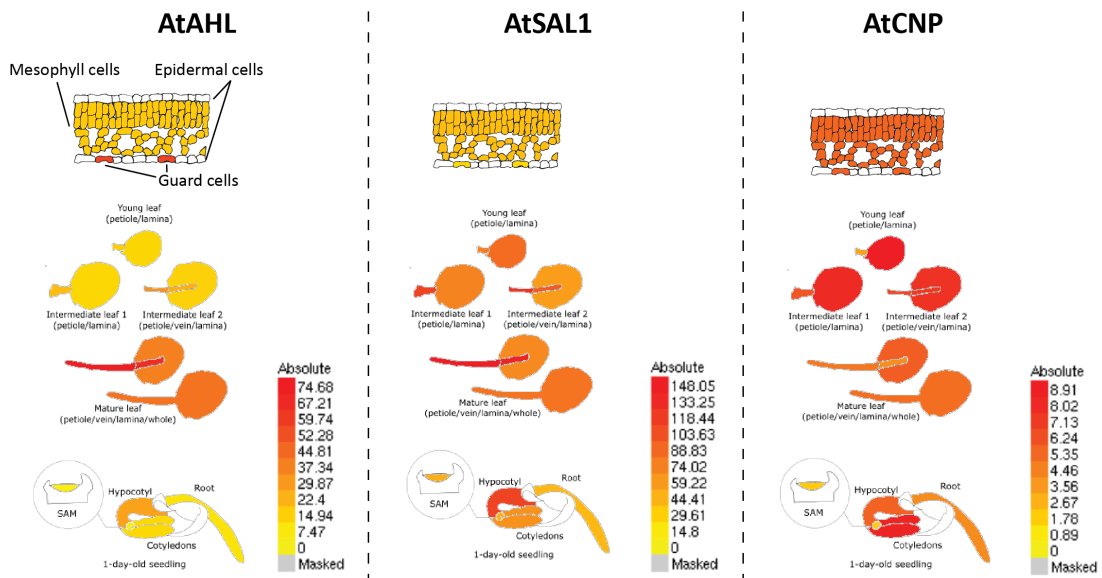


Figure 4.7: AtAHL, AtSAL1, and AtCNP tissue expression: top, leaf cross section; middle, leaves in different stages of development; bottom, 1 day old seedling. Expression is relative with each paralogue. Expression is shown in a gradient from red showing highest expression, yellow shows low expression. Images are reproduced from TAIR, data originally from Klepikova et al. (2016).

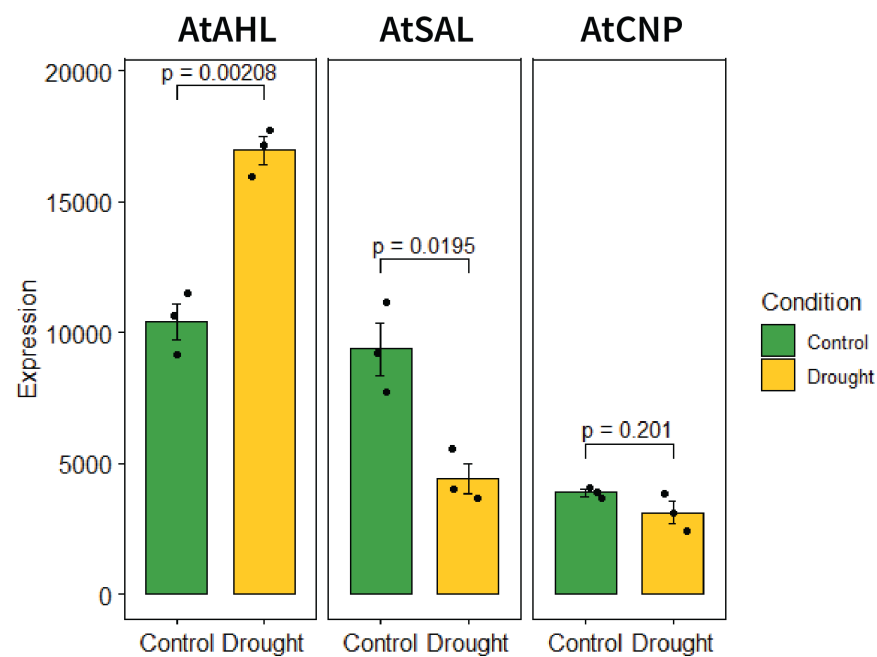


Figure 4.8: AtSAL1, AtAHL, AtCNP expression (mRNA) under drought compared to the control with the p-value. Dataset is from GENEVESTIGATOR, experiment ID AT-00626. Rosette leaf samples of WT Col-0 grown for 30 days on solerite soil without water limitation under 16 h light ($120 \mu\text{mol photons m}^{-2}\text{s}^{-1}$) / 8 h dark cycles at 22 °C. Transcriptome profiling array: Affymetrix *A. thaliana* ATH1 Genome Array was used to provide relative mRNA abundance for the three genes. Statistics (Welch *t*-test, Cohen's *d* effect size), error bars show standard error - AtAHL: Statistic = -7.59, degrees of freedom = 3.75, effect size = -6.20 (large), AtSAL1: Statistic = 4.33, degrees of freedom = 3.23, effect size = 3.54 (moderate), AtCNP: Statistic = 1.76, degrees of freedom = 2.35, effect size = 1.4 (negligible).

Table 4.4: Significant results* from AtSAL1, AtAHL, and AtCNP co-expressed gene ontology analysis

Gene	GO analysis	Detail	Fold enrichment	p-value
AtSAL1	Cellular component	Mitochondrial ATP synthase complex (GO:0000275)	58.43	5.67×10^{-4}
		Chloroplast stroma (GO:0009570)	11.83	3.33×10^{-30}
	Biological process	S-glycoside biosynthetic process (GO:0016144)	>100	2.55×10^{-35}
		Glucosinolate biosynthetic process (GO:0019761)	>100	2.55×10^{-35}
		Tryptophan biosynthetic process (GO:0000162)	>100	2.15×10^{-8}
	Molecular function	Adenylylsulfate kinase activity (GO:0004020)	>100	1.73×10^{-8}
		Desulfoglucosinolate sulfotransferase activity (GO:0047364)	>100	8.01×10^{-7}
		Tryptophan synthase activity (GO:0004834)	>100	1.38×10^{-6}
AtAHL	Cellular component	NS		
	Biological process	S-glycoside biosynthetic process (GO:0016144)	>100	7.45×10^{-28}
		Glucosinolate biosynthetic process (GO:0019761)	>100	7.45×10^{-28}
		Glycosyl compound biosynthetic process (GO:1901659)	>100	2.38×10^{-25}
	Molecular function	Adenylylsulfate kinase activity (GO:0004020)	>100	2.61×10^{-9}
		Adenylyl-sulfate reductase (glutathione) activity (GO:0033741)	>100	3.60×10^{-8}
		Adenylyl-sulfate reductase activity (GO:0009973)	>100	3.60×10^{-8}
AtCNP	Cellular component	Chloroplast thylakoid lumen (GO:0009543)	53.7	4.64×10^{-18}
	Biological process	Photosystem II complex assembly (GO:0010270)	>100	1.83×10^{-5}
		Nitrogen utilization (GO:0019740)	>100	7.29×10^{-5}
		Ammonia assimilation cycle (GO:0019676)	>100	9.91×10^{-5}
	Molecular function	Peptidyl-prolyl cis-trans isomerase activity (GO:0003755)	32.16	9.91×10^{-6}

*See Methods, NS = none significant. Top GO terms for the co-expressed genes with AtSAL1, AtAHL, and AtCNP. Fold enrichment for those terms are shown along with the p-values.

4.4 Discussion

4.4.1 Technical findings

The SSN approach proved effective for protein functional classification, enabling identification of orthology relationships and discovery of previously uncharacterised protein families. By mapping sequence similarity networks, this method successfully distinguished three functionally distinct nucleotide phosphatase clusters (SAL, AHL, and the novel CNP) that would be difficult to separate using traditional pairwise sequence comparisons alone. The network-based classification provided initial functional hypotheses that could be validated through subsequent analyses.

Subcellular localisation prediction using DeepLoc2 (Thummuluri et al., 2022) was crucial for functional classification of the unclassified CNP clade. The predominant chloroplast localisation prediction, combined with co-expression analysis showing association with thylakoid lumen enzymes, enabled confident functional assignment despite the absence of wet-lab characterisation. This demonstrates how computational localisation prediction can guide functional classification of uncharacterised protein families. The 69.2% plastid localisation prediction requires careful interpretation. DeepLoc2 achieves a Matthews Correlation Coefficient (MCC) of approximately 0.90 for plastid localisation (Thummuluri et al., 2022; Armenteros et al., 2019), but MCC quantifies population-level classifier performance rather than per-prediction confidence. Because the false negative rate (plastid proteins incorrectly predicted as non-plastid) is not reported, positive predictions carry greater interpretive weight than negative ones. The observed 69.2% could reflect fully conserved plastid targeting if the remainder represent false negatives, although this interpretation remains speculative without quantitative false negative rate data.

PANTHER co-expression and gene ontology analysis (Mi et al., 2017) provided additional layers of functional prediction. The correlation between CNP and photosystem II-associated enzymes, combined with tissue-specific expression patterns, strengthened the functional classification. Similarly, AHL co-expression with sulphur metabolism enzymes suggested functional similarity to SAL despite sequence divergence (Table 4.4).

The phylogenetic approach enabled functional prediction through evolutionary context. Short branch lengths between clades indicated functional similarity, and duplication events suggested neo- or sub-functionalisation. Statistical analysis of sequence properties (length, aromaticity, isoelectric point) between outparalogous clades provided quantitative support for functional di-

vergence predictions in **AHL** (Table 4.1). The **CNP** outparalogous clades showed no significant differences in these properties (Table 4.2), and combined with branch support below 95%, this suggests the apparent clade split represents a phylogenetic artefact rather than functional divergence.

The wheat SAL analysis validated this classification framework through independent evidence. The **SSN** approach distinguished wheat SAL sequences into functionally distinct groups that corresponded to phylogenetic clades with greater than 95% bootstrap support. These classifications aligned directly with opposing phenotypic outcomes in field trials, validating the methodology's ability to predict functional divergence from sequence relationships. The co-expression network analysis provided independent support demonstrating distinct enrichment patterns, with **TaSAL1** associated with core metabolic processes and **TaSAL2** with stress-responsive pathways, correctly predicting the direction of yield effects upon deletion. This supports enhancing functional prediction accuracy by integrating multiple lines of computational evidence before conducting *in plantae* experiments.

4.4.2 Biological findings

The phylogenetic classification revealed three functionally distinct nucleotide phosphatase families with specific cellular roles. The novel **CNP** which I've classified as a chloroplast-localised enzyme associated with photosystem II in developing tissues (Table 4.4 and Figure 4.7) hints at a previously unrecognised function in photosynthetic regulation. Its close relationship to SAL, which acts as a retrograde stress sensor (Chan et al., 2016), suggests **CNP** may also function in stress regulation. One untested hypothesis is that **CNP** shifts metabolic flux away from photosystem II biosynthesis under stress, conserving resources during growth. Alternatively, it may fine-tune retrograde **PAP** (or a **PAP**-analogue) signalling during development.

The phylogeny is of high statistical quality, generated from 400 replicates with typically well-supported branch values. Two notable limitations warrant discussion.

The first concerns the apparent **CNP** duplication, which most likely represents a phylogenetic artefact rather than genuine functional divergence, as discussed above.

The second limitation concerns the divergence point between SAL and **AHL**, which has poor branch support. The inferred topology places this divergence such that SAL contains only streptophyte sequences, and the **AHL** clade includes both chlorophytes (Algal **AHL**) and streptophytes (Basal **AHL**, **AHL- α** ,

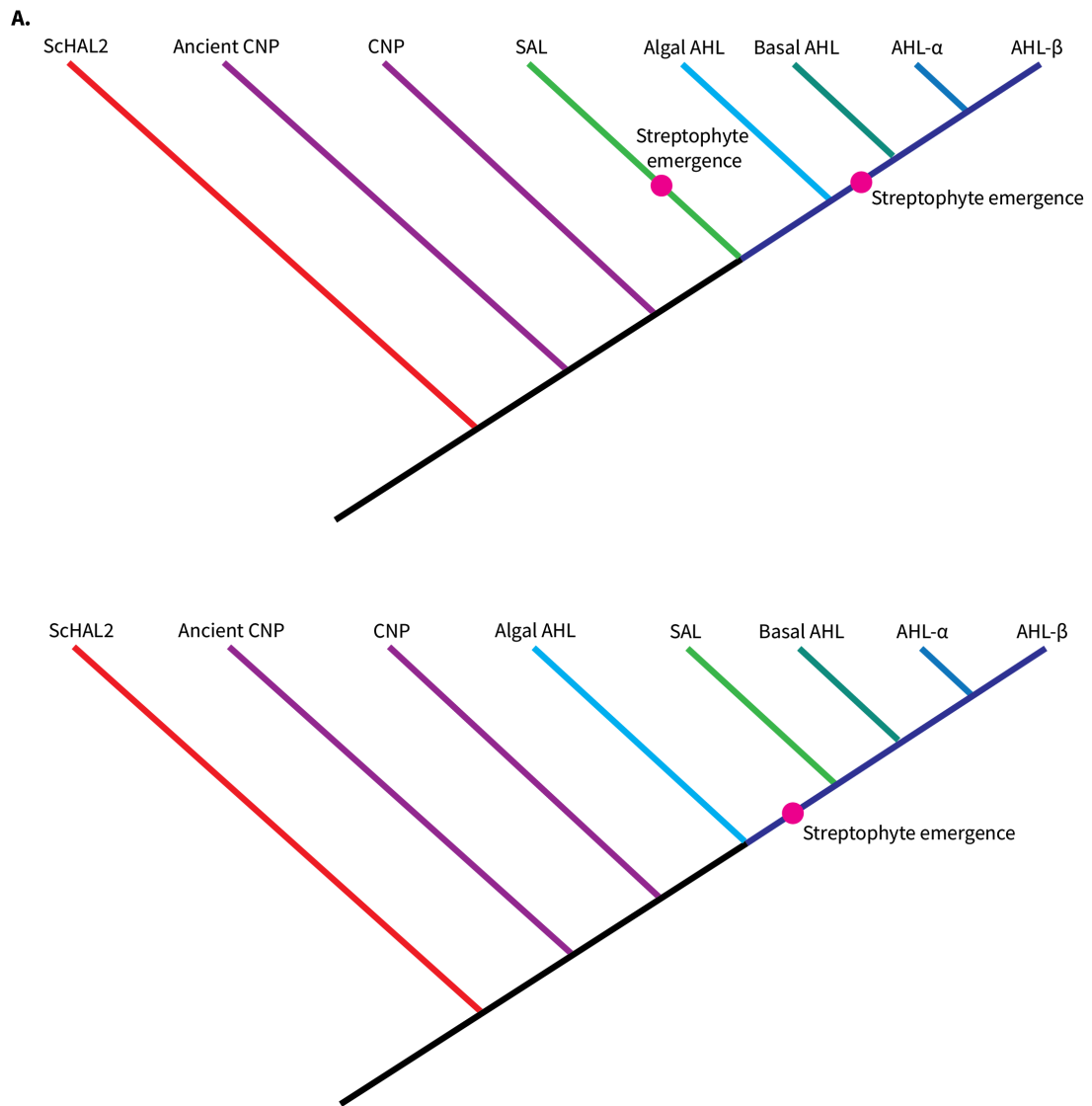


Figure 4.9: Cladogram illustrating alternative topologies for the SAL–AHL divergence. **A.** The inferred topology, in which the duplication producing SAL and AHL occurred before the chlorophyte–streptophyte divergence. This scenario requires subsequent loss of SAL from chlorophytes. Pink circles indicate streptophyte emergence, which under this topology must be marked at two points on the gene tree. **B.** An alternative topology in which chlorophytes retained the ancestral gene as Algal AHL, with the duplication producing SAL and land plant AHL occurring subsequently within streptophytes. The poor branch support at the SAL–AHL node means neither topology can be confidently favoured based on sequence data alone.

AHL-β). Two contrasting scenarios are presented to illustrate how alternative branch placements would affect interpretation of SAL–AHL evolutionary history (Figure 4.9).

Under the inferred topology (Figure 4.9A), the gene duplication producing SAL and AHL occurred before the chlorophyte–streptophyte divergence. This scenario requires that chlorophytes subsequently lost SAL while retaining AHL. If this loss occurred once at the base of chlorophytes, it represents a single evolutionary event and is not inherently implausible.

An alternative topology (Figure 4.9B) would place SAL as sister to land

plant **AHL** lineages (Basal **AHL**, **AHL- α** , **AHL- β**) rather than to the entire **AHL** clade. Under this scenario, chlorophytes retained the ancestral gene as Algal **AHL**, and a subsequent duplication within streptophytes produced **SAL** and land plant **AHL** as distinct paralogues. Chlorophytes would never have possessed **SAL** because the duplication had not yet occurred.

Distinguishing between these scenarios requires adequate sampling of chlorophyte sequences. The chlorophyte representation in this dataset derives almost entirely from Chlorophyceae, with families such as Volvocaceae and Haematococcaceae sampled multiple times, and major chlorophyte classes including Trebouxiophyceae, Ulvophyceae, and prasinophyte-grade lineages are absent or poorly represented. This sampling limitation reflects the sparse sequencing of chlorophyte genomes in public databases rather than methodological choices in this study. Consequently, the absence of chlorophyte **SAL** sequences cannot be confidently attributed to genuine absence versus incomplete sampling.

Resolving the phylogenetic topology definitively would require broader chlorophyte sampling as additional genomes become available. However, the current dataset permits an alternative approach through functional analysis. Sequence-based phylogenetic methods cannot resolve this uncertainty because they assume functional similarity correlates with sequence similarity. Comparison of active site properties may provide independent evidence bearing on the evolutionary relationships among these paralogues, though interpretation requires integrating both sequence and functional data rather than treating either as definitive. Chapter 5 addresses this question through direct comparison of active site electrostatic fields, providing an independent line of evidence for the functional relationships among these paralogues, but cannot conclusively elucidate the true evolutionary path of **PAP** catabolism.

Given the poor branch support at this node and inadequate chlorophyte sampling, the precise timing of **SAL**–**AHL** divergence remains unresolved by sequence-based phylogenetic analysis alone. This uncertainty concerns the deep evolutionary history of the gene family and does not affect the core conclusions of this chapter. The identification of **CNP** as a distinct chloroplast-localised family rests on an independent, well-supported clade with consistent localisation prediction, co-expression, and tissue expression evidence. The statistical differentiation between **AHL- α** and **AHL- β** concerns relationships within land plant **AHL**, where branch support is robust. Similarly, the wheat field trial validation concerns functional divergence between **TaSAL1** and **TaSAL2** within the **SAL** clade, entirely independent of when **SAL** diverged from **AHL**. The unresolved question is specifically whether Algal **AHL** represents a lineage that lost **SAL** or one that predates the **SAL**–land plant **AHL** duplication. Chapter 5 examines

the functional relationships among these paralogues, providing an independent perspective that complements the sequence-based analysis.

Functional classification through evolutionary patterns identified SAL under purifying selection, with the absence of retained duplicates from ancestral polyploidy events indicating gene dosage sensitivity (Figure 4.3). The implications of this dosage sensitivity for polyploid wheat SAL genes are examined in Section 4.4.4.

The AHL family was functionally classified as drought-responsive based on multiple prediction criteria. Despite sequence similarity to SAL (indicated by short branch lengths in Figure 4.3), AHL shows distinct functional specialisation in *A. thaliana* through primarily guard cell expression (Figure 4.7), increased drought-responsive expression (Figure 4.8), and PAP catalytic activity (Gil-Mascarell et al., 1999). The retention of duplicate AHL copies following the γ -polyploidy event, combined with statistical differentiation between the outparalogous clades (Table 4.1), strongly suggests neo- or sub-functionalisation (including tissue or developmental sub-functionalisation). This contrasts with SAL, where duplication is not well tolerated, indicating different functional constraints between these related families.

The evolutionary distribution of PAP-degrading enzymes reveals a functional gap in chlorophyte algae. Yeast ScHAL2 and land plant SAL both degrade PAP and share high sequence similarity (Figure 4.1), yet chlorophyte algae possess AHL but lack SAL. This suggests ancestral AHL fulfilled the PAP degradation role before the gene duplication event that created SAL, which occurred coincident with land plant evolution (Figure 4.3). Following duplication, SAL assumed PAP degradation function and developed drought-sensing coincident with plant adaptation to land, and AHL underwent neo- or sub-functionalisation. The PAP catalytic activity retained by modern AHL in *A. thaliana* has significantly impaired binding specificity (K_M of 160 μM compared to 2–10 μM for AtSAL1 (Gil-Mascarell et al., 1999)), likely representing vestigial ancestral function or tissue-specific activity. Guard cell localisation of AHL in *A. thaliana* theoretically supports specialisation for stomatal regulation under drought conditions, although this requires experimental testing.

The functional classification approach revealed that *A. thaliana*, and likely also Brassicaceae more broadly, appears to have lost the AHL- β paralogue, which is unusual compared to other core eudicots (Figure 4.3). Without this phylogenetic functional classification, research efforts might be misdirected by studying an evolutionary outlier. This finding emphasises the value of taxonomically broad functional classification before detailed mechanistic studies, and suggests that computational approaches with large sample sizes can ad-

dress classification questions across broad taxonomic groupings more effectively than studies limited to single model organisms.

The phylogenetic framework also explains why previous wheat studies produced inconsistent results (Manmathan et al., 2013; Abdallah et al., 2025; Mohr et al., 2022). By targeting all SAL copies simultaneously, these approaches disrupted both core metabolic functions (TaSAL1) and stress-responsive pathways (TaSAL2), creating unpredictable phenotypes dependent on experimental conditions. The evolutionary analysis revealed these studies were conflating functionally distinct genes. The neo-functionalisation model, where TaSAL1 maintained ancestral functions in core metabolism and TaSAL2 evolved specialised stress-responsive roles, accounts for the 10-fold expression difference and distinct co-expression networks observed between these groups.

The combination of neo-functionalisation and dosage sensitivity provides a framework for targeted crop improvement, with the parallel evolution of SAL duplications across grass subfamilies supporting potential transferability to other polyploid cereals.

4.4.3 *In vivo* validation of sequence-based classification

The field trials across 15 diverse environments provided robust *in vivo* validation of the phylogenetic predictions. The opposing yield effects, where TaSAL2 deletions improved performance by 4–8% and TaSAL1 deletions reduced it, directly confirmed the functional divergence predicted by evolutionary analysis. This represents empirical validation that phylogenetic classification can predict phenotypic outcomes in complex polyploid genomes.

The severe yield penalty in double mutants (18% reduction) demonstrated that both homologous groups serve essential but distinct functions, validating the prediction of neo-functionalisation rather than redundancy. The consistency of these effects across diverse field environments spanning different water availability and temperature regimes confirms that the phylogenetically predicted functions are maintained under variable conditions.

The water productivity improvements in TaSAL2 deletion lines from the field, coupled with yield increases, validated the co-expression network prediction of specialised stress-responsive functions. This *in plantae* evidence demonstrates that evolutionary and computational analyses can identify targets for crop improvement that enhance performance without fitness penalties.

Importantly, the field validation confirms the classification *method*, not merely the biological conclusion about wheat SAL. The logical chain runs from phylogenetic classification predicting functional divergence, to expression and co-

expression data supporting distinct roles, to field trials confirming opposing phenotypic effects. Each step built on the previous, with the final validation demonstrating that sequence-based classification, when applied with appropriate evolutionary context, can generate actionable predictions for polyploid crop improvement.

4.4.4 Gene dosage sensitivity in SAL evolution

The evolutionary pattern of SAL across Viridiplantae presents an apparent paradox. Phylogenetic analysis reveals strong purifying selection with no retained duplicates from multiple major ancestral polyploidy events including ζ at spermatophyte origin, ϵ at angiosperm origin, and γ at core eudicot divergence (Figure 4.3). This pattern indicates that SAL is dosage-sensitive and that duplicate copies are eliminated following whole-genome duplication. Yet hexaploid wheat possesses multiple SAL gene copies across two distinct paralogue groups. Given this we need to ask how a gene family that consistently loses duplicates persist as multiple copies in wheat?

The answer also lies in distinguishing between different modes of gene duplication. The wheat SAL copies did not arise from whole-genome duplication. Instead, they trace to a Pooideae-specific gene duplication that occurred before the speciation of barley (*Hordeum vulgare*), *Brachypodium distachyon*, and *Aegilops tauschii*. All diploid progenitors of wheat already carried both **TaSAL1** and **TaSAL2** paralogue groups before the hybridisation events that produced hexaploid bread wheat. Understanding why this particular duplication persisted, while ancient whole-genome duplications consistently lost SAL copies, requires examining how different duplication modes interact with dosage constraints.

The gene balance hypothesis provides a framework for understanding differential retention of duplicates from different origins (Birchler and Veitia, 2012). Whole-genome duplication doubles all genes simultaneously, preserving the stoichiometric relationships among interacting proteins. This coordinated doubling maintains cellular homeostasis because relative gene dosage remains constant even as absolute copy number increases. In contrast, small-scale duplications such as tandem or segmental duplications increase the copy number of individual genes or small gene sets while leaving their interaction partners unchanged. This creates stoichiometric imbalance that can disrupt protein complex assembly, signalling cascade kinetics, and metabolic flux (Freeling and Thomas, 2006).

Dosage-sensitive genes therefore show contrasting fates depending on du-

plication mode (Qiao et al., 2019). Following whole-genome duplication, dosage-sensitive genes initially remain as duplicates because losing one copy would create the same imbalance that small-scale duplication causes. Over time, compensatory mechanisms including expression divergence and protein sequence evolution gradually release paralogues from these constraints, eventually permitting loss of duplicate copies through pseudogenisation or deletion (Clark and Donoghue, 2018). The exponential decay of duplicate retention observed across plant lineages reflects this gradual release from dosage balance requirements.

SAL follows this expected pattern. No retained duplicates from the ancient ζ , ϵ , or γ whole-genome duplications appear in the phylogeny, consistent with the duplicate loss expected for dosage-sensitive genes over evolutionary time. The absence of retained SAL duplicates from these ancient events indicates that any duplicates arising from whole-genome duplication were subsequently eliminated, as predicted for genes under dosage constraints.

The Pooideae-specific duplication that produced *TaSAL1* and *TaSAL2* represents a different situation. This duplication is evolutionarily recent, occurring after the divergence of Pooideae from other grass subfamilies but before the speciation of barley and wheat progenitors. Recent duplications have not yet had time for the gradual compensatory drift that eventually permits duplicate loss. Under the gene balance hypothesis, such duplications should be actively selected against unless they provide immediate fitness benefits that outweigh the costs of dosage imbalance.

The retention of both *TaSAL1* and *TaSAL2* across Pooideae indicates that this duplication was stabilised by rapid functional divergence rather than lost like earlier whole-genome duplication copies.

Several lines of evidence support neo-functionalisation of the *TaSAL1* / *TaSAL2* duplication. Expression analysis reveals 10-fold higher expression of *TaSAL1* compared to *TaSAL2* across developmental stages (Figure 4.6B). Co-expression network analysis shows enrichment for distinct functional categories where *TaSAL1* associates with fundamental cellular processes including sulphur metabolism, DNA repair, and aminoacyl-tRNA pathways, and *TaSAL2* associates with metabolic flexibility pathways including carbon metabolism and amino acid metabolism (Figure 4.6C). The chapter discussion interprets this pattern as *TaSAL1* maintaining ancestral core metabolic functions and *TaSAL2* evolving specialised roles enabling metabolic adjustment under variable conditions. Most directly, field trials demonstrated opposing fitness effects where *TaSAL2* deletion improved yield by 4–8% and *TaSAL1* deletion reduced yield, with double mutants showing severe 18% yield penalties (Table 4.3).

This functional divergence resolves the apparent paradox of SAL copy number in wheat. Before neo-functionalisation, duplicate SAL copies would have been subject to the same dosage constraint that eliminated copies from ancient whole-genome duplications. The evolution of distinct roles meant that *TaSAL1* and *TaSAL2* no longer occupied the same functional niche, with each paralogue participating in different pathways with different stoichiometric requirements.

Critically, neo-functionalisation did not simply release dosage constraints. Rather, it transformed a single constraint against redundant SAL duplicates into two distinct, paralogue-specific constraints reflecting each gene's new functional role. *TaSAL1*, integrated into core cellular processes, requires sufficient expression to maintain these functions and shows fitness costs when deleted. *TaSAL2*, with its distinct metabolic associations, shows fitness benefits when copy number is reduced, suggesting that polyploid copy number may exceed optimal levels for its specialised role. The functional divergence thus enabled retention of both paralogues by creating distinct selective regimes that no longer conflict.

The parallel but independent evolution of SAL duplications in both Pooideae and Panicoideae subfamilies (Figure 4.6A) suggests that functional divergence may represent a recurring evolutionary response to SAL gene duplication in grasses. This convergent pattern indicates that selective pressures favouring SAL functional divergence operate broadly across Poaceae, though the poor branch support within Panicoideae prevents confident interpretation of duplication timing and functional consequences in that subfamily. The conservation of both *TaSAL1* and *TaSAL2* across multiple Pooideae species confirms that functional divergence stabilised this duplication before the radiation of the subfamily.

The polyploidisation history of bread wheat provides context for understanding current SAL copy numbers. Modern hexaploid wheat (*Triticum aestivum*, genome AABBDD) arose through two hybridisation events. The first occurred approximately 800,000 years ago when *Triticum urartu* (genome AA) hybridised with a species related to *Aegilops speltoides* (genome BB), producing tetraploid emmer wheat (genome AABB). The second hybridisation approximately 8,500 years ago joined this tetraploid with *Aegilops tauschii* (genome DD), yielding modern hexaploid bread wheat (Marcussen et al., 2014).

This history is important for understanding SAL dosage. The Pooideae-specific duplication that produced *TaSAL1* and *TaSAL2* predates wheat progenitor speciation. Both *Hordeum vulgare* (barley) and *Aegilops tauschii* retain both paralogue groups, and since barley diverged from the Triticeae lineage before the speciation events that gave rise to wheat progenitors, the dupli-

tion must have occurred earlier still. Hexaploidisation therefore did not create the *TaSAL1*/*TaSAL2* distinction but rather multiplied pre-existing paralogue groups across subgenomes.

The resulting copy number in hexaploid wheat reflects this history. The Chara variety possesses six SAL gene copies. Specifically three *TaSAL1* copies on chromosomes 5A, 5B, and 5D, and three *TaSAL2* copies on chromosomes 4A, 7A, and 7D (7D was missing from the transcript databases likely due to using an older scaffold, with the 4A localisation resulting from a known 4AL/7BS translocation (Devos et al., 1995). The extra *TaSAL1* copies being due to transcript variants that were retained in the analysis).

The yield improvement observed upon *TaSAL2* deletion raises the question of whether polyploid SAL dosage exceeds optimal levels. If hexaploidisation multiplied ancestral diploid copy numbers across subgenomes, the resulting dosage may exceed that which evolved under purifying selection in diploid ancestors. From this perspective, targeted deletion could return the system toward a more optimal dose rather than simply removing a functional gene.

Several observations support this interpretation. The purifying selection observed across SAL evolution (Figure 4.3) indicates that ancestral diploid dosage was maintained within narrow optimal ranges. The absence of retained duplicates from the ζ , ϵ , and γ whole-genome duplications demonstrates that SAL resists dosage increases. The field trial results show that *TaSAL2* deletion produces consistent yield improvements of 4–8% across 15 environments (Table 4.3), suggesting that the deletion addresses a genuine dosage excess rather than a condition-specific effect.

Under this model, the *TaSAL2* deletion phenotype reflects optimisation rather than loss of function. The neo-functionalisation of *TaSAL2* toward metabolic flexibility pathways, evidenced by co-expression with carbon and amino acid metabolism genes (Figure 4.6C), may create a system where responsiveness scales with gene copy number. In a diploid ancestor, two *TaSAL2* copies would provide appropriate metabolic adjustment capacity. In a hexaploid with three or more copies, this same responsive system may over-react to metabolic signals, imposing costs through excessive or premature activation of stress responses. Reducing copy number would restore appropriate responsiveness.

There are alternative explanation to consider. The yield improvement could reflect disruption of a net-negative regulatory pathway rather than dosage optimisation. Under this model, *TaSAL2* participates in signalling cascades that limit growth under non-stressed conditions, and its removal permits constitutive high growth. This differs from dosage optimisation in implying that *TaSAL2* function is inherently yield-limiting rather than over-represented. Distinguish-

ing these models would require measuring *TaSAL2* expression or PAP substrate levels across knockout lines to determine whether the phenotype correlates with total SAL activity or with complete removal of a specific function.

A second alternative is that *TaSAL2* deletion permits compensatory up-regulation of *TaSAL1* or other genes that provide the growth benefit. Under this model, *TaSAL2* itself is neutral or mildly positive, but its presence suppresses a more beneficial alternative. Expression analysis comparing *TaSAL1* levels in wild-type versus *TaSAL2* deletion backgrounds would test this hypothesis.

Finally, the yield improvement might reflect escape from a pathogen susceptibility locus linked to *TaSAL2* rather than effects of SAL function itself. The consistency of effects across 15 environments spanning different pathogen pressures makes this explanation less likely, but it cannot be excluded without pathology-controlled experiments.

Despite these alternatives, the dosage optimisation model provides the most parsimonious explanation consistent with the evolutionary evidence. The pattern of purifying selection, the absence of retained duplicates from ancient polyploidy events, and the contrasting effects of *TaSAL1* versus *TaSAL2* deletion all support the interpretation that SAL dosage operates under constraint and that polyploid *TaSAL2* copy number exceeds optimal levels.

The opposing yield effects of *TaSAL1* and *TaSAL2* deletions demonstrate that functional divergence created distinct dosage requirements for each paralogue. This differential sensitivity extends beyond simple retention or loss to reveal qualitatively different relationships between copy number and fitness.

For *TaSAL1*, deletion consistently reduced yield (Table 4.3), indicating that full polyploid dosage is required for optimal performance. The associations from the co-expression network analysis with fundamental cellular processes, including sulphur metabolism, DNA repair, and aminoacyl-tRNA pathways (Figure 4.6C), suggest that *TaSAL1* participates in core housekeeping functions that scale directly with cellular activity. In a polyploid with multiple cell lineages each requiring active metabolism, the three subgenomic *TaSAL1* copies may collectively provide necessary enzyme activity. The 10-fold higher expression of *TaSAL1* compared to *TaSAL2* (Figure 4.6B) supports this interpretation. High constitutive expression would be expected for a gene whose product is required in stoichiometric quantities for core cellular processes.

For *TaSAL2*, deletion improved yield by 4–8%, indicating that reduced dosage provides a fitness advantage. The association with metabolic flexibility pathways suggests *TaSAL2* functions in responsive adjustment rather than constitutive activity. If *TaSAL2* modulates metabolism in response to stress signals, excess copies could create an over-sensitive system that imposes costs through

inappropriate activation. The lower and more variable expression pattern of *TaSAL2* across developmental stages is consistent with a regulatory role that responds to conditions rather than maintaining constant activity.

This contrast illuminates how neo-functionalisation transforms dosage constraints. Before functional divergence, both paralogue copies would have been subject to the same constraint against SAL dosage increases. The evolution of distinct functions partitioned this single constraint, with *TaSAL1* now requiring sufficient copies to maintain core metabolic flux and *TaSAL2* requiring limited copies to maintain appropriate stress responsiveness. Neither constraint existed before the duplication because the ancestral gene served either both functions within a single dosage optimum.

Opposite-sign single-mutant yields demonstrate that *TaSAL1* and *TaSAL2* are not interchangeable (4.3). The 18% double-mutant penalty exceeds the expected multiplicative result from the single mutant, indicating a synergistic interaction where loss of both paralogues compounds deleteriously despite their opposing single-mutant effects.

This non-redundancy validates the neo-functionalisation interpretation and explains why both paralogue groups persisted across Pooideae evolution despite originating from a dosage-sensitive gene family.

It is important to note that the double mutant (*SAL1.1* × *SAL2.1*) represents deletion of one copy from each paralogue group (*TaSAL1-5D* and *TaSAL2-4A*), not complete elimination of both groups. The severe phenotype from this partial reduction underscores that both paralogue groups serve essential functions even when most copies remain intact. This also suggests the remaining copies cannot fully compensate for the lost ones, further supporting distinct functional roles.

The framework of paralogue-specific dosage optima has implications for crop improvement in polyploid species beyond wheat. Polyploidisation may systematically create dosage imbalances for genes that have undergone neo-functionalisation, with each paralogue potentially exceeding or falling short of its specific optimum. Identifying which paralogues would benefit from dosage adjustment requires understanding both the functional divergence between paralogues and the relationship between copy number and function for each specialised role. The SAL example demonstrates that genes with regulatory or responsive functions may be more amenable to dosage reduction than those with housekeeping or structural roles.

The dosage optimisation model and its alternatives generate distinct predictions that could be tested through targeted experiments. Three approaches would most effectively discriminate between competing explanations for the

TaSAL2 deletion phenotype.

Expression-based discrimination. The dosage optimisation model predicts that yield improvement correlates with total SAL activity rather than complete removal of a specific function. Measuring PAP substrate levels across a TaSAL2 copy number series would test this directly. If PAP accumulation shows a linear relationship with copy number reduction, this would support dosage optimisation. Conversely, if PAP levels change abruptly only upon complete TaSAL2 removal, this would favour the regulatory pathway disruption model. Similarly, if TaSAL1 expression increases in TaSAL2 deletion backgrounds, this would support the compensatory up-regulation hypothesis. RNA-seq comparing wild-type and TaSAL2 deletion lines could quantify any such compensation.

Dosage series experiments. The models make different predictions about the relationship between TaSAL2 copy number and yield. Dosage optimisation predicts a graded response where intermediate copy numbers should produce intermediate phenotypes. Lines with one remaining TaSAL2 copy (as opposed to complete deletion) should show yield improvements between wild-type and null mutants. Creating an allelic series through CRISPR targeting of individual TaSAL2 copies would distinguish these predictions.

Cross-variety and cross-species comparisons. Natural variation in SAL copy number across wheat germplasm provides opportunities to test dosage predictions without genetic manipulation. Surveying TaSAL2 copy number across diverse wheat accessions and correlating with yield performance under controlled conditions would reveal whether reduced copy number associates with improved performance more broadly. The prediction extends to other polyploid crops in Pooideae, where barley, with its single copy of each paralogue group, should represent the ancestral dosage optimum. If the dosage model is correct, introducing additional TaSAL2 copies into barley should impose yield costs proportional to the copy number increase.

Further refinements. Testing partial TaSAL2 reduction through an allelic series would determine whether yield improvements scale with copy number reduction, as the dosage model predicts, or appear only upon complete deletion. Systematic epistasis analysis examining whether TaSAL2 deletion benefits depend on specific TaSAL1 alleles or copy numbers would clarify whether the two paralogue groups can be optimised independently in breeding programs.

These predictions extend the methodology developed in this thesis beyond classification toward quantitative genetic analysis. Where the phylogenetic and electrostatic approaches identified which genes to target, dosage analysis would refine how much to modulate each target. This progression from clas-

sification to optimisation represents a natural development of sequence-based functional prediction toward practical crop improvement.

4.4.5 Limitations of sequence-based approaches

The classification methods employed here have characteristic limitations that provide the scope of questions they can address. Notably, the Welch *t*-tests comparing sequence properties between outparalogous clades do not account for phylogenetic non-independence. Phylogenetically corrected methods such as phylogenetic generalised least squares or phylogenetic ANOVA would be more appropriate for comparative data and may yield different conclusions, particularly for marginal cases.

More fundamentally, sequence-based classification operates down at the sequence-structure level of abstraction. Sequence similarity provides an imperfect proxy for functional similarity because two enzymes with high sequence identity might have identical catalytic properties or might differ substantially, depending on where differing residues are located and how they affect the active site environment. The SSN approach could identify that AHL- α and AHL- β arose from gene duplication and showed statistical differences in sequence properties, but could not determine whether these differences translate to catalytic divergence or merely reflect neutral drift with functional conservation.

The nucleotide phosphatase analysis explicitly recognised this limitation, given that catalytic functional divergence is defined by the active site and can be approximated by the shape of the binding pocket and its electrostatic environment (Figure 4.4).

Chapter 5 develops this approach using Fourier decomposition of electrostatic potentials calculated from protein structures. Applied to the same nucleotide phosphatase dataset analysed here, electrostatic comparison successfully distinguished SAL, AHL, and CNP as functionally distinct groups. The method also revealed other function-dependent clustering patterns. These findings will demonstrate that electrostatic analysis can detect functional differences invisible to sequence-based methods, addressing the gap identified in the current chapter.

4.4.6 Testable hypotheses

The functional predictions derived from this phylogenetic classification approach require experimental validation but provide a framework for targeted investigation. The purifying selection identified for SAL predicts that species with recent SAL duplications should show negligible expression of extra copies, con-

sistent with rapid pseudogenisation. Artificially modulating SAL copy number in crop species with recent duplications could potentially improve growth or stress tolerance, providing an alternative hypothesis for the yield improvements observed in wheat SAL knockout lines beyond the neo-functionalisation model.

The **CNP** classification as a chloroplast-localised, photosystem II-associated enzyme generates predictions testable through localisation studies, expression analysis under photosynthetic stress, and phenotypic characterisation of knockout lines. The **AHL** functional divergence model predicts that the β -paralogue retains ancestral **AHL** functionality in species that possess it, noting that *A. thaliana* appears to have lost **AHL**- β . This model also predicts that chlorophyte **AHL** and land plant SAL should cluster in functional similarity networks but diverge in sequence-based analyses, reflecting distinct evolutionary trajectories from a common ancestral **PAP**-degrading function.

The gene dosage sensitivity predicted for SAL and the neo-functionalisation of **AHL** provide complementary mechanisms that together explain the evolutionary patterns observed across the nucleotide phosphatase family. These testable predictions, generated from computational and evolutionary analysis, would not be readily apparent from **wet-lab** experiments alone. The field trial validation of wheat SAL classification demonstrates that such predictions can have practical agricultural applications when translated to polyploid crop species.

Chapter 5

Enzyme active site electrostatic field comparison and similarity network

5.0.1 How this chapter is structured

This chapter focused on the development technology to classify enzymes through electrostatic similarity. The Methods section in this chapter details established protocols unique in this thesis to this chapter, such as statistical methods. The Results section covers the development and application of the protocols. The Discussion section ties together the Discussions from Chapters 3 and 4 as a foundation for understanding and contextualising the results from this chapter. It then discussed the biological interpretation of the results from this chapter, how this novel method relates to the existing methods for protein classification.

5.0.2 Significance statement

- **Methodological Approach:** Here we validate a Fourier decomposition approach with Sinkhorn optimal transport for comparing enzyme active site electrostatic fields. By decomposing three-dimensional electrostatic maps into frequency bands corresponding to different spatial scales, the protocol isolates biologically informative signal from noise arising from structural modelling uncertainty and numerical artefacts. Mid-frequency bands (spatial wave-lengths of 3.7–6.7 Å) corresponding to secondary structure scales provided the strongest discriminatory power, explaining up to 33% of variance in paralogue classification across 506 enzymes at five pH conditions.
- **Key Discovery:** Electrostatic comparison revealed functional relationships invisible to sequence and structural analysis. Algal AHL and Ancient CNP from Chromista, Rhodophyta, and Glaucophyta clustered electro-

statically with SAL despite phylogenetic classification as distinct paralogues. This electrostatic similarity suggests these enzymes share catalytic properties with SAL, potentially filling the **PAP** catabolic function in lineages where SAL is absent. pH sensitivity analysis revealed regulatory divergence between **AHL** sub-clades: **AHL- α** showed strong pH-dependent electrostatic clustering while **AHL- β** showed none, indicating functional specialisation following gene duplication that is undetectable by sequence comparison.

- **Biological Significance:** The electrostatic approach extends the sequence-structure-function paradigm by directly comparing the physical properties governing enzymatic catalysis. The unexpected clustering of Algal **AHL** and Ancient **CNP** with SAL suggests a continuous lineage of **PAP** catabolic capacity across Viridiplantae, resolving an apparent functional gap spanning hundreds of millions of years of eukaryotic evolution. The differential pH sensitivity between **AHL- α** and **AHL- β** provides evidence for regulatory neo-functionalisation following gene duplication, offering a mechanism explaining long-term retention of duplicate copies with minimal sequence divergence.

5.1 Introduction

The preceding chapters of this thesis demonstrate a systematic progression in computational approaches for protein functional classification, each revealing fundamental constraints and increasing methodological sophistication. Chapter 3 showed using established techniques that **AF2**-generated structural models could elucidate mechanisms of enzyme dysfunction through analysis of phytoene synthase mutations. While addressing the research questions the structural analysis failed to capture electrostatic relationships governing both substrate- and protein-protein interactions. Chapters 3 also employed **Evolocity** network analysis to classify **DXO1** and **eIF4E** homologues across thousands of sequences, successfully identifying angiosperm-specific innovations while remaining constrained by assumptions linking sequence similarity to functional conservation. Chapter 4 integrated **SSNs** with phylogenetic inference to resolve nucleotide phosphatases into three distinct families (**SAL**, **AHL**, **CNP**) and predicted functional divergence following duplication events in **AHL**. This supports my assertion that these gene-copies from duplication events in plant orders should not be discounted as simple redundant copies but rather neo- or sub-functionalisation should be the null hypothesis. Subsequently, Chapter 4 also provided field validation of this assertion, demonstrating that **TaSAL2** deletion enhanced wheat

yield while **TaSAL1** deletion diminished yield. We do not know how these SAL copies functionally relate to other SALs.

This chapter addresses a critical gap in the structure-function relationship, specifically that it is the electrostatic environment that directly mediates catalytic activity. Through integration of high-throughput structural modelling with systematic comparison of electrostatic fields, this approach goes beyond inference-based methods to directly quantify functional differences between homologous enzymes. The nucleotide phosphatase family, characterised by variable evolutionary diversity, agricultural significance, and minimal classification identified in Chapter 4, provides an excellent system for validating this methodology.

5.1.1 Why Electrostatics

The canonical sequence-structure-function paradigm in structural biology requires refinement specifically when considering enzymes. While some protein classes may have their function less explicitly defined by electrostatics, such as structural proteins, the core function of enzymes is catalysis which is directly defined by electrostatics and the spatial distribution within steric constraints. Enzymatic function is mediated only indirectly by three-dimensional structure. The electrostatic environment is more directly causal to defining catalytic function. The electrostatic environment comprises the spatial distribution of electric charge that governs catalytic mechanisms. This electrostatic field represents a more fundamental level of functional abstraction, integrating all sequence and structural factors that ultimately drive enzymatic catalysis.

Electrostatic fields orchestrate enzyme function at several levels. They govern substrate recognition and binding through complementary charge distributions that generate electric gradients. Like the banks of a river channelling water, these gradients direct molecular trajectories. For example, phosphorylated substrates, such as **PAP**, can be electrostatically guided towards regions of positive potential and hydrophobic molecules are excluded from charged domains. Following substrate binding, the precise spatial arrangement of charges within the active site near the enzyme surface selectively stabilises reaction transition states, thereby reducing activation energies. The charge distributions additionally facilitate proton transfer reactions between catalytic residues and substrates, with the local electric field modulating pKa values and consequently determining protonation states at physiological pH.

Unlike sequence or structural approaches, protein electrostatics inherently captures environmental modulation; a critical factor in functional classification

as enzymes do not exist orthogonally to their environment. Environmental pH exemplifies this principle, where changes in pH fundamentally restructure the electrostatic field through altered ionisation states of titratable residues. An aspartate residue functioning as a nucleophile at pH 7.0 may become protonated and catalytically inactive at pH 6.5. Similarly, a histidine residue serving as a general base at pH 7.6 may lose this capacity at pH 8.2. These pH-dependent modifications to active site charge distribution explain the characteristic narrow pH optima observed for most enzymes. Enzymes must maintain specific electrostatic environments essential for catalysis beyond protein stability considerations, and the environment affects this.

Beyond the immediate active site environment, long-range electrostatic interactions are fundamental to protein architecture and molecular recognition processes. Complementary electrostatic surfaces between proteins facilitate complex formation (McCoy et al., 1997). The spatial distribution of surface charges additionally influences protein stability through networks of salt bridges and hydrogen bonds that maintain tertiary structure. Perturbations to these electrostatic networks, whether through mutation or post-translational modification, can propagate allosteric conformational changes that modulate enzymatic activity in the active site. Given the importance of the electrostatic environment governing enzyme function, methods have been developed to analyse it by first calculating an electrostatic map.

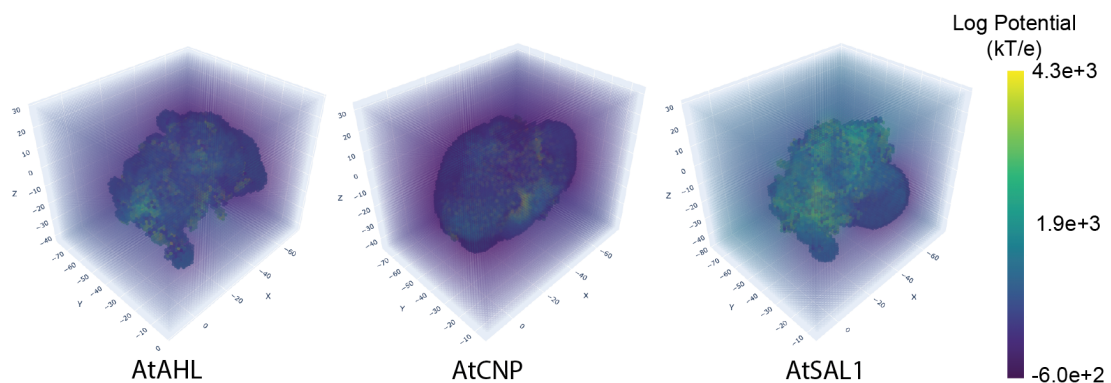


Figure 5.1: Representative electrostatic map images. From left to right, AtAHL, AtCNP, and AtSAL1. Voxels coloured by charge log potential (kT/e). Axis units are Å. Environmental voxels have been reduced in display size to allow viewing of the central protein.

An electrostatic map constitutes a three-dimensional representation of the electric potential field surrounding a protein molecule, calculated from the spatial coordinates and partial charges of all constituent atoms and environmental influences. An electrostatic map is a three-dimensional grid of charge potentials through space, where each point in the grid is assigned the value of the charge potential at that location. This is analogous to a two-dimensional image

where each point (a pixel) carries a value representing a specific colour. Pixels in a three-dimensional image are called voxels. An electrostatic map can be viewed as a three-dimensional image where charge points are represented as voxels coloured according to their charge potential (Figure 5.1). These maps delineate regions of positive electrostatic potential, where cations are repulsed and anion species attracted, and regions of negative potential exhibiting opposite characteristics. Figure 5.1 shows the electrostatic maps of AtAHL, AtCNP and AtSAL1 at pH 7.9. Importantly, not only is charge varied within the protein (the opaque mass the centre of each image), but we can see the varied impact of the protein on the surrounding electrostatic environment. In contrast to sequence alignments or structural superpositions, electrostatic maps directly represent the physical environment encountered by ligands and other interacting elements. The interacting elements interact not only at the protein surface but with the electrostatic field in the surrounding environment. These representations integrate the collective contributions of all charged residues, backbone dipole moments, bound cofactors, ions, and ordered water molecules into a unified functional descriptor of the enzyme's function. That is, every functional element of an enzyme is encoded in the electrostatic map; including but not limited to kinetics, substrate specificity, stability under different conditions, pH optima, conformational dynamics, and post-translational regulation. By limiting ourselves to just the sequence-structure-function paradigm we gloss over the most important determinant of enzyme function.

5.1.2 Technical Advances Open Up Opportunity

Given that electrostatic fields encode enzymatic function, direct comparison of these fields should provide superior functional discrimination compared to sequence or structural similarity metrics alone. Better functional discrimination explicitly results in better classification. Enzymes exhibiting substantial sequence divergence may nevertheless maintain virtually identical electrostatic features within their active sites, thereby preserving function despite evolutionary distance. Conversely, single amino acid substitutions producing minimal structural perturbation may dramatically alter local electric fields, fundamentally modifying substrate specificity or catalytic mechanism. Given this, functional classification using electrostatic comparisons should be superior to sequence- or structural-homology based approaches.

There has been limited adoption of electrostatic comparison methodologies. This reflects multiple technical constraints that only recently are starting to be overcome, including difficulties in determining protein structural models and

computational challenges in comparing electrostatic maps. Of these, limited protein structures has been the main constraint. Without a structure the electrostatic map cannot be calculated. Protein structure determination via crystallography or NMR spectroscopy typically require months to years per protein, rendering comprehensive family-wide structural characterisation infeasible.

Protein structure prediction

Prior to AF2, protein structure prediction methods could not achieve the accuracy required for reliable electrostatic comparisons. The biennial [Critical Assessment of protein Structure Prediction \(CASP\)](#) competition provides standardised benchmarks for the field. At CASP13 in 2018, the best non-AlphaFold methods achieved average GDT_TS scores of 50–55 for challenging targets. Using established conversion relationships where GDT_TS values of 90 and 80 correspond to [Root-Mean-Square Deviation \(RMSD\)](#) values of approximately 1.5 Å and 2.5 Å respectively ([Kryshtafovych et al., 2021](#)), these scores translate to typical errors of 5.0–6.0 Å. Given that the AtSAL1 active site spans less than 15 Å, such errors are substantial before considering the sub-angstrom positioning requirements of catalytic residues. Template-based approaches like I-TASSER, MODELLER, and RosettaCM performed reasonably when templates shared greater than 30 % sequence identity with the target, but accuracy deteriorated rapidly below 25 % identity where suitable templates are often unavailable for plant proteins ([Lin et al., 2025](#); [Rasheed et al., 2020](#)). Early deep learning methods including trRosetta improved contact prediction but still achieved only moderate accuracy for most targets ([Kryshtafovych et al., 2019](#)). Without high-throughput accurate structural modelling, electrostatic classification of large homologue sets was not practical.

AF2 demonstrated a leap-forward advancement at CASP14, achieving a median GDT_TS of 92.4 across all targets compared ([Jumper et al., 2021](#)). For *de novo* modelling targets, historically the most challenging category, AF2 achieved a median GDT_TS of 87 while the best alternative methods struggled to exceed 60 ([Kryshtafovych et al., 2021](#)). A GDT_TS of 90 indicates that predictions capture not only correct fold topology but also accurate side-chain conformations and loop regions. This near-atomic accuracy, recognised by the 2024 Nobel Prize awarded to Demis Hassabis and John Jumper, enables quantitative analysis of active site geometry and catalytic residue positioning. The system's [pLDDT](#) scores provide per-residue confidence estimates that correlate strongly with actual model accuracy, enabling identification of reliable regions for downstream applications ([Jumper et al., 2021](#)).

AF2 does have limitations with oligomeric complexes and some NMR-derived structures (Kryshtafovych et al., 2021), but these are not relevant to this study as I examine only monomeric nucleotide phosphatases. Validation of my AF2 models against the AtSAL1 crystal structure confirmed excellent performance for this enzyme family. Where previous methods provided approximate folds suitable for hypothesis generation, AF2 delivers the accuracy required for electrostatic map comparison. This capability made prediction of the 506 structures essential for this study feasible within days using research-accessible computational resources, as demonstrated in Chapter 3.

Electrostatic calculation methods

Several computational methods exist for calculating electrostatic potentials around proteins, each with distinct trade-offs between accuracy and computational cost.

Molecular dynamics with explicit solvent

Molecular dynamics simulations with explicit solvent molecules provide the most detailed treatment of electrostatic interactions. The cost to this is that they require extensive conformational sampling over nanosecond to microsecond timescales, requiring days or weeks of computation per protein, often on supercomputers (Klepeis et al., 2009; Jones et al., 2022). For comparative electrostatics across homologues rather than conformational dynamics, this additional computational expense is not justified and would be computationally intractable for large datasets.

Coulombic calculations

At the opposite extreme, Coulombic calculations apply Coulomb's law ($V = k \frac{q}{r}$) to treat ionised residues as point charges in a vacuum. These calculations can generate electrostatic potential maps by calculating the potential at each point on a grid surrounding the protein. However, vacuum-based calculations drastically underestimate the importance of solvent screening and ionic shielding because they ignore water molecules and mobile ions that attenuate electrostatic interactions in solution (Onufriev and Case, 2019). In cellular environments, the high dielectric constant of water and the presence of dissolved ions create a screening effect that reduces the effective range and magnitude of electrostatic forces. The Debye length, the characteristic distance over which electrostatic interactions decay due to ionic screening, is approximately 8 Å at physiological ionic strength (150 mM). Beyond this distance, mobile counter-

ions effectively shield charges from one another, rendering long-range electrostatic interactions negligible under physiological conditions (Israelachvili, 2011).

Generalised Born models

Generalised Born (GB) models represent a computationally efficient approximation to the Poisson-Boltzmann equation. These models treat the solvent as a dielectric continuum, eliminating the need to explicitly represent individual solvent molecules while retaining essential electrostatic screening effects. The approach requires calculation of position-dependent Born radii (R_i) for each atom, which quantify the degree of solvent exposure and the effective distance from each atomic charge to the solvent dielectric boundary. Surface-exposed atoms exhibit Born radii approaching their van der Waals radii, while buried atoms display larger effective Born radii reflecting reduced solvent accessibility (Onufriev and Case, 2019).

While GB calculations run faster than full Poisson-Boltzmann solutions, making them attractive for high-throughput screening, they sacrifice accuracy in geometrically complex regions. Deep binding pockets, narrow channels, and highly charged surfaces — features prevalent in enzyme active sites — pose particular challenges for GB approximations (Onufriev and Case, 2019).

Empirical pK_a prediction

Empirical pK_a prediction tools, including PROPKA and H++, offer a computationally efficient alternative to physics-based approaches (Olsson et al., 2011; Anandakrishnan et al., 2012). PROPKA uses a rule-based algorithm that calculates pK_a shifts by summing empirical contributions from local structural features including hydrogen bonding patterns, desolvation effects, and electrostatic interactions with neighbouring charged groups. H++ takes a hybrid approach, combining Poisson-Boltzmann calculations with empirical corrections derived from training data.

Both methods achieve substantial computational efficiency by avoiding iterative solutions of the Poisson-Boltzmann equation across all possible protonation states. However, training data dependency imposes fundamental constraints on accuracy and generalisability. Empirical methods show reduced reliability when encountering molecular features poorly represented in their training sets, including any non-standard amino acids, post-translational modifications, unusual metal coordination environments, and proteins with highly charged surfaces or buried salt bridges. Plant proteins from plant-unique sub-cellular environments are likely under-represented in training sets. Physics-

based approaches derive predictions from fundamental electrostatic principles, offering consistent treatment of novel environments provided the underlying physical model remains valid (Onufriev and Case, 2019; Baker et al., 2001).

Poisson-Boltzmann approach

The Poisson-Boltzmann approach addresses the limitations of the above methods while maintaining computational feasibility. The method numerically solves the Poisson-Boltzmann equation, which describes electrostatic potential distributions around biomolecules in solution by combining Poisson's equation (relating charge density to electrostatic potential) with the Boltzmann distribution of mobile ions. The physics-based approach extrapolates reliably to broad datasets, making it particularly suitable for studying pH-dependent phenomena across diverse proteins (Baker et al., 2001).

Adaptive Poisson-Boltzmann Solver (APBS) solves this equation using finite difference methods on a three-dimensional grid. The calculation accounts for the low dielectric environment inside the protein (typically $\epsilon = 2-4$), the high dielectric solvent ($\epsilon = 78.5$ for water), and the screening effects of mobile ions at physiological ionic strength (150 mM). The resulting electrostatic potential maps provide quantitative values at each grid point, which molecular visualisation tools like PyMOL (Schrödinger LLC. (2020)) can render as colour-coded surfaces.

APBS requires preprocessing of the structural files using PDB to PQR converter) (PDB2PQR). PDB2PQR serves as a preprocessing tool that bridges the gap between crystallographic structures or structural models and APBS's electrostatic calculations. The software parses standard Protein Data Bank (PDB) protein structure files and converts them to PQR format, a modified PDB format that includes atomic charge (Q) and radius (R) parameters essential for electrostatic calculations. During this conversion, PDB2PQR performs several critical preprocessing steps. The software adds missing hydrogen atoms based on optimal protonation states, since hydrogen positions are not modelled by AF2 and are not resolved in crystal structures. It then optimises the hydrogen bonding network through systematic rotation of hydroxyl and thiol groups, sampling different orientations to minimise the overall energy. For charge parameterisation, force fields such as AMBER assign partial charges derived from quantum mechanical calculations and have been extensively validated for protein electrostatics (Cornell et al., 1996). The software calculates theoretical pK_a values for all ionisable residues using a modified Tanford-Kirkwood method, which accounts for desolvation effects and electrostatic interactions between titratable groups. The Tanford-Kirkwood approach treats the protein as a low dielectric

medium embedded in a high dielectric solvent, modelling ionisable residues as point charges and solving for electrostatic interactions using continuum electrostatics theory. This method explicitly considers desolvation effects. This is the energetic penalty incurred when an ionisable group buried within the protein interior loses its hydration shell. Since water (dielectric constant ~ 80) stabilises charged species far more effectively than the protein interior (dielectric constant $\sim 4\text{--}20$), burial of an ionisable group shifts its pK_a value. The carboxyl groups become less acidic (higher pK_a) and amino groups become less basic (lower pK_a) when removed from the aqueous environment. These pK_a predictions determine the protonation state of each residue at the specified pH, directly influencing the overall charge distribution of the protein.

Electrostatic map comparison methods

Electrostatic fields guide substrate binding and dictate catalytic features. For enzymes within the same family, comparing these electrostatic patterns could reveal how evolution has fine-tuned catalytic function despite structural divergence. While well-established protocols exist for sequence alignment and structural superposition, no standardised methodologies exist for quantifying similarity between electrostatic maps of non-identical structures.

This gap limits the utility of protein structures for classification. We can calculate detailed electrostatic maps yet we lack robust methods to compare them systematically across different proteins, particularly at scale. Most existing methods work well when examining the same protein under different conditions, comparing how pH changes or mutations alter the electrostatic landscape of a single structure (Blomberg et al., 1999). But when proteins have different structures, even closely related homologues, direct comparison breaks down. A charged residue in one protein might occupy a completely different position in three-dimensional space compared to its evolutionary counterpart, making point-by-point comparison meaningless. The core challenge in comparing electrostatic maps lies in establishing spatial correspondence between different protein structures.

Grid-based comparison methods

Comparison is straightforward when proteins share identical structures. The structures are aligned (superimposed) and differences calculated between electrostatic values at corresponding grid points. The RMSD method exemplifies this approach (Blomberg et al., 1999). RMSD calculates the square root of the averaged squared differences between electrostatic potentials at each point,

providing a single numerical value representing overall dissimilarity. For identical proteins under different conditions, **RMSD** provides an intuitive and quantitative measure of electrostatic changes (Blomberg et al., 1999).

However, **RMSD** fails when applied to non-identical proteins. Even homologous enzymes with conserved function often display structural variations including loops of different lengths, shifts in secondary structure positioning, or mutations within the active site. Each structural difference means that grid points no longer correspond between proteins. **RMSD** treats these as completely different features rather than recognising them as the same functional element in slightly different positions. **RMSD** also provides no normalisation for proteins with different total charges: a protein with twice the net charge automatically shows higher **RMSD** values regardless of whether the spatial distribution patterns are functionally equivalent.

Recognising these limitations, researchers developed correlation-based indices that partially address the normalisation problem. The Hodgkin similarity index calculates correlation coefficients between electrostatic potentials (Hodgkin and Richards, 1987). The related Carbo index employs similar correlation mathematics but uses a different normalisation scheme that constrains values between 0 and 1 (Carbó et al., 1980). By sacrificing the ability to detect anti-correlation, the Carbo index provides more intuitive interpretation for comparing proteins expected to have similar functions. However, the Carbo index suffers from sensitivity to outliers in the electrostatic distribution and can be dominated by regions of very high potential that may not be functionally relevant (Good, 1992). Both indices perform reasonably well for closely related structures where most regions can be meaningfully paired after superposition, yet they degrade rapidly with increasing structural divergence. Unaligned regions such as loops, insertions, or structural variations that cannot be matched between proteins contribute noise that eventually overwhelms any signal from conserved electrostatic features.

Regional and surface-focused approaches

Given the limitations of whole-protein comparison, several methods have been developed that focus on specific regions where electrostatic properties are most functionally relevant. **Protein Interaction Property Similarity Analysis (PIPSA)** pioneered this approach by restricting comparison to the protein surface where molecular recognition occurs (Wade et al., 2001; Blomberg et al., 1999; Richter et al., 2008). The method identifies a “skin” region extending approximately 3 Å from the protein surface, representing the catalytically important interaction region. **PIPSA** computes electrostatic potentials on a regular three-dimen-

sional grid surrounding each protein, then identifies grid points within this specified distance from any surface atom. Only these near-surface points contribute to similarity calculations using either Hodg-kin or Carbo indices.

PIPSA handles structural variation through a multi-step alignment process. Proteins first undergo structural superposition using backbone atoms to establish the best overall alignment. The method then applies a distance-based filter that excludes grid points lacking corresponding surface regions in both proteins. If one protein has an inserted loop creating additional surface area, only grid points with surface correspondence in both structures contribute to the similarity calculation. This selective inclusion means that structural variations are neither penalised nor rewarded but simply excluded from analysis. The approach works well when conserved surface regions carry the functionally important electrostatic features but can miss important differences when insertions or deletions alter functional sites.

This surface focus sidesteps many structural variation problems. Internal structural differences become irrelevant if they do not affect surface electrostatics. Alternative packings of the hydrophobic core, different arrangements of buried secondary structures, or variations in domain orientations only matter if they change the surface charge distribution. The method has successfully classified protein families with extreme sequence divergence, including cytochrome P450 enzymes sharing less than 20 % sequence identity (Werck-Reichhart and Feyereisen, 2000).

However, PIPSA's surface restriction creates blind spots for certain protein functions. Many enzymes contain deep binding pockets where substrates must navigate through channels to reach buried active sites. In these cases, electrostatic gradients extend well beyond 3 Å from the surface, creating "electrostatic funnels" that guide substrate approach and orientation. These internal electrostatic features, critical for catalytic function, remain invisible to PIPSA's surface-restricted analysis. The method also assumes that the protein surface can be clearly defined, which becomes problematic for proteins with large cavities or multiple interconnected channels.

More sophisticated regional approaches emerged to address these limitations. MatchTope focuses specifically on peptide-binding grooves in Major Histocompatibility Complex I (MHC) proteins, defining cylindrical regions based on crystallographic peptide positions (Mendes et al., 2022). The method typically examines cylinders extending 40 Å radius by 2 Å length, encompassing the entire binding interface. By dividing this cylinder into smaller sub-volumes and calculating Hodgkin indices within each, MatchTope generates detailed similarity maps of binding sites. This geometric approach successfully predicts

cross-reactivity between MHC molecules with highly divergent sequences, demonstrating that functional similarity can be captured through careful regional definition.

Alignment-independent transformation methods

Some modern comparison methods seek to avoid the alignment problem altogether through approaches that transform the comparison into a different mathematical space. Spherical harmonic decomposition projects electrostatic potentials onto spherical surfaces centred at the protein's centre of mass (Ritchie and Kemp, 2000; Morris et al., 2005). This transformation works by sampling electrostatic potentials at regular angular intervals on concentric spherical shells, typically ranging from the protein surface to 10–15 Å beyond. Each shell's potential distribution undergoes decomposition into spherical harmonic basis functions, which are mathematical functions that form a complete set for describing any pattern on a sphere.

This decomposition is somewhat analogous to Fourier decomposition, which is discussed later in this chapter. Low-order spherical harmonic terms capture broad charge distributions, revealing whether the protein is generally positive on one face and negative on another. High-order terms encode fine details including the precise positions of individual charged residues or the detailed shape of binding pocket electrostatics. By discarding phase information that depends on molecular orientation, the method achieves rotation invariance. Two proteins can be compared regardless of how they are oriented in space, eliminating the need for structural alignment entirely.

This mathematical elegance comes with biological trade-offs. The spherical projection assumes proteins are roughly globular and that their centre of mass provides a meaningful reference point. These assumptions fail for many protein architectures. Elongated proteins like fibronectin, horseshoe-shaped proteins like leucine-rich repeat domains, or proteins with multiple flexibly connected domains all map poorly onto spherical surfaces. A horseshoe-shaped protein with its active site in the concave interior would have this functionally critical region poorly represented when projected onto spheres centred at its geometric centre. The dimensionality reduction inherent in spherical projection also discards the complex three-dimensional structure of active sites, potentially obscuring functionally critical local features that do not project cleanly onto spherical surfaces.

Modern Computational Capabilities

The transformation of electrostatic map comparison from computational im-

possibility to routine analysis reflects broader changes in computational biology. Contemporary [Graphical Processing Units \(GPUs\)](#) provide orders-of-magnitude acceleration for Poisson-Boltzmann calculations and on a current mid-range commercial-off-the-shelf [GPU](#) we can calculate hundreds of electrostatic map comparisons simultaneously through massive parallelisation. Software packages have developed from command-line tools requiring specialised expertise to automated pipelines that have graphical interfaces and sensible default parameters. The combination of [AF2](#)'s comprehensive structural predictions ([Jumper et al., 2021](#)) and modern computational infrastructure means that systematic electrostatic analysis of entire protein families now becomes feasible with computing power available to many research groups. This capability was impossible just a decade ago. Despite these advances, comparison of complete protein electrostatic maps remains computationally demanding. Three-dimensional grids encompassing entire protein structures contain millions of data points, and exhaustive pairwise comparison still pushes computational limits. Practical solutions involve intelligent restriction of comparison regions. This study sought to make millions of comparisons and by focusing on the active site I hoped to extract functional "active site fingerprints". This targeted approach makes computational requirements manageable while retaining the most functionally relevant information.

5.1.3 Improving nucleotide phosphatase classification

The nucleotide phosphatase family highlights the advantage of using electrostatic analysis to augment traditional classification approaches. Chapter 4 identified unexpected functional diversity within this enzyme family with three closely related but distinct phylogenetic paralogues ([AHL](#), [CNP](#), [SAL](#)) exhibiting unclear functional delineation, evidence for neo/sub-functionalisation following gene duplication events, and complex subcellular localisation predictions. The [AHL](#) paralogues demonstrated statistical evidence of functional divergence through differential sequence lengths, aromaticity values, and isoelectric points; however, the functional consequences of this divergence remained uncharacterised. Without mechanistic understanding of how sequence variations translate to functional modifications, identification of valuable engineering targets remains speculative, as does explanation for the unique loss of [AHL- \$\beta\$](#) in *Arabidopsis* amongst eudicots raised in chapter 4.

The agricultural relevance of nucleotide phosphatases also necessitate improved functional classification. Chapter 4 demonstrated that targeted modulation of [SAL](#) expression in wheat enhanced yield by 4-8% through selective

deletion of *TaSAL2* while preserving *TaSAL1* function. This discovery emerged from phylogenetic classification but required extensive multi-environment field trials for validation. Should electrostatic analysis successfully predict functional outcomes directly from structural data, crop improvement could be substantially accelerated through identification of beneficial modifications prior to field testing.

From an evolutionary perspective, nucleotide phosphatases exhibit greater complexity than may be typical in *Animalia*. Nucleotide phosphatases have prokaryotic origins (Kerk et al., 2021) with subsequent divergent evolution in plant lineages. This evolutionary divergence generated distinct paralogues at key evolutionary points including land adaption. Representative characterisation of these paralogues requires analytical methods capable of detecting functional modifications not apparent at the sequence level or structural level. Given that function can diverge rapidly as sequence similarity decrease, using a single ($n=1$) statistically representative sequence from each paralogue or even several for analysis is insufficient, rather a representative sample set, with n defined by power analysis (Nakagawa and Cuthill, 2007), is required and this is beyond what is currently practical for *wet-lab* experimentation. This problem requires a computational approach where trading single sequence classification precision for clade-wide representative precision. That is, *wet-lab* experiments offer precision but not broad classification accuracy. Conversely these computational approaches provide broad classification accuracy at the expense of functional precision for any one sequence.

The predominance of *A. thaliana*-focused studies has created substantial knowledge gaps. Most functional characterisation has concentrated on *AtSAL1*, yet as demonstrated in chapter 4 *A. thaliana* represents an atypical system having lost *AHL- β* where as other eudicots retained both paralogues. This sampling bias suggests current understanding may derive from an evolutionary outlier rather than a representative system. Comprehensive functional classification across plant phylogenetic diversity would reveal far more of the functional space occupied by nucleotide phosphatases.

Plant genome complexity substantially complicates functional classification. Frequent polyploidy events generate multiple gene copies that undergo rapid neo- or sub-functionalisation. Chapter 4 illustrated this complexity as wheat contains six *SAL* copies with opposing functional effects, which traditional sequence-based classification approaches failed to distinguish. Classification methodologies must therefore operate at scales commensurate with genomic diversity, rather than the family-wide classifications that are common.

The extensive natural diversity within plant nucleotide phosphatases re-

mains largely unexploited. Database repositories contain thousands of sequences representing potential sources of stress tolerance alleles, yield enhancement traits, or novel catalytic activities. Without robust functional classification, this diversity cannot be systematically evaluated or utilised. Electrostatic comparison could identify functionally equivalent enzymes across taxonomic boundaries, facilitating trait transfer between crop species, while deepening our understanding of molecular evolution.

The combination of evolutionary significance, agricultural applications, and technical feasibility positions nucleotide phosphatases as a suitable system for validating electrostatic-based functional classification. Successful demonstration with this enzyme family would establish the methodology's applicability to other systems where sequence-based approaches inadequately capture functional classification. Systematic comparison of active site electrostatics across evolutionary space could represent a fundamental advancement in the prediction and classification of enzymes.

5.1.4 Hypothesis

Plant nucleotide phosphatase paralogues cluster according to active site electrostatic similarity, while the within-paralogue distances are significantly smaller than between-paralogue distances (tested via PERMANOVA). Inferred by this (but not tested yet) is that proteins that cluster together based on their active site electrostatic environments should have more similar function (specifically catalytic activity).

Aims:

1. Develop a computational pipeline for comparing enzyme electrostatic properties
2. Test whether paralogue classification can be achieved via electrostatic similarity
3. Validate the approach against sequence/structural comparison methods

5.1.5 Data source

Computational approaches addressing biological questions at scales beyond experimental feasibility have established precedent in leading journals. Thiltgen et al. (2023) used simulated data and revealed protein evolutionary dynamics across $10^{4.2}$ generations using atomistic simulation which is impossible to recreate experimentally. Qiu et al. (2021) demonstrated global fitness optima by computationally screening 480 sequences from 160,000 variants, replacing

years of laboratory work. My nucleotide phosphatase analysis presents an even more extreme case. Using 506 structures at five pH values requires 3,199,185 pairwise electrostatic comparisons. This computational approach enables comparative analysis across 506 structures at five pH conditions, which is a scale that is infeasible through experimental characterisation alone, while generating testable predictions for targeted validation. Moreover, Chapter 4 demonstrated that *A. thaliana* uniquely lost AHL- β among eudicots, illustrating how single-organism studies miss evolutionary patterns. Instead, this work analyses real protein sequences from established databases, representing thousands of naturally occurring variants. The electrostatic classifications undergo rigorous statistical validation including, but not limited to, hierarchical clustering, **Permutational Multivariate Analysis of Variance (PERMANOVA)** to measure statistical significance and effect size, cluster cohesion analysis, and permutation analysis. This approach trades single-enzyme precision for statistically robust patterns across entire protein families without relying on simulated data. This approach is essential to address evolutionary questions spanning kingdoms rather than individual proteins, while demonstrating real-world tractability.

5.2 Methods

5.2.1 Computer system

All calculations, unless otherwise stated, were performed on the following computer system:

- NVIDIA RTX 3080 10 GB memory
- AMD Ryzen Threadripper 3960X 24-Core Processor 2200MHz
- 256 GB memory (random access memory (RAM))

5.2.2 Code availability

All coding was done in Python (3.10). The electrostatic comparison pipelines and analysis code are available in Jupyter Notebooks at github:

https://github.com/mattmorts-sci/PhD_code

5.2.3 Structural modelling

Prior to structural modelling, sequence transit peptides were removed manually. All sequences were aligned through **MSA** using **MAFFT** with default parameters. AtSAL1 has an experimentally determined transit peptide with a well

conserved cut site at a 'MAYE' amino acid motif. The [MSA](#) was manually curated in Geneious 2023.2.1 to ensure this motif was correctly aligned and using the AtSAL1 as a reference all residues upstream of this motif were removed.

Structure modelling used Colabfold Batch version 1.4.0 with the [AF2](#) model on the National Computational Infrastructure (NCI) GADI supercomputer. Default parameters were used with the exception of the following which were specified. The **search process** was in database load mode 1, using 12 threads. Resources used per structure were 12 [Central Processing Units \(CPUs\)](#), max memory was 48 GB, max wall-time was 3 hours. **Structure prediction** process used AMBER force field to relax structure using [GPU](#) with 3 recycles. Templates were not used. Resources used per structure were 12 [CPUs](#) and 1 [GPU](#), max memory was 48 GB, max wall-time was 3 hours. Structure's prediction were treated as failed where resources were exceeded. Failed structures were run a second time on GADI. Those that continued to fail were re-run locally with extended wall-time of 6 hours. Those that failed that were dropped from the analysis.

Structures were aligned pairwise against AtSAL1 using the open-source version of PyMol version 3.1.0, with the Python code in Appendix B PyMol structural alignment [B.1](#).

5.2.4 Electrostatic calculations

Electrostatic calculation parameters

Electrostatics were calculated locally using PDB2PQR version 3.7.1 and [APBS](#) version 3.0. [PDB2PQR](#) calculations used the AMBER force field. [APBS](#) was run with the solvent modelled with a dielectric constant of 78.4 (representing bulk water), while the protein interior was assigned a dielectric constant of 2. The molecular surface was calculated using a solvent probe radius of 1.5 Å with a bulk solvent density of 10.0 points/Å². Surface smoothing was applied using a cubic spline window of 0.3 Å. All calculations were performed at 298.15 K. Charges were discretised onto the grid using cubic B-spline interpolation. The apolar contribution to solvation free energy was calculated using a cavity surface tension of 0.0251 kcal/mol/Å².

Calculations were performed at pH values of 7.0, 7.3, 7.6, 7.9, and 8.2, corresponding to biologically relevant pH levels for plant subcellular compartments:

- plastid stroma light (pH 8.0–8.2) ([Trinh and Masuda, 2022](#); [Heldt et al., 1973](#))
- plastid stroma dark (pH 7.0) ([Trinh and Masuda, 2022](#))

- cytoplasm (pH 7.0–7.2, or 7.4–7.8 for chlorophytes) (Shen et al., 2013; Küsel et al., 1990)
- guard cell cytoplasm (pH 7.5) (Blatt and Armstrong, 1993; Grabov and Blatt, 1999)
- mitochondrial matrix (pH 7.8–8.0) (Shen et al., 2013)
- peroxisome (pH 8.2) (Shen et al., 2013)

This generated 2,530 electrostatic maps total (506 structures $\times$ 5 pH values).

Cropping electrostatic maps

The electrostatic maps were cropped to the active site programmatically creating a rectangular prism of 16.0 $\times$ 18.5 $\times$ 20.0 Å (D $\times$ W $\times$ H). The cropping coordinates were determined by manually defining the bottom plane of a rectangular prism to encompass the active site of the experimentally determined AtSAL1 (Figure 5.2). This was checked to ensure that the volume encompassed the inferred (through structural homology) active sites of *A. thaliana* AHL, and CNP. A random selection of 10% cropped structure were checked to ensure the inferred active sites were contained in the volume.

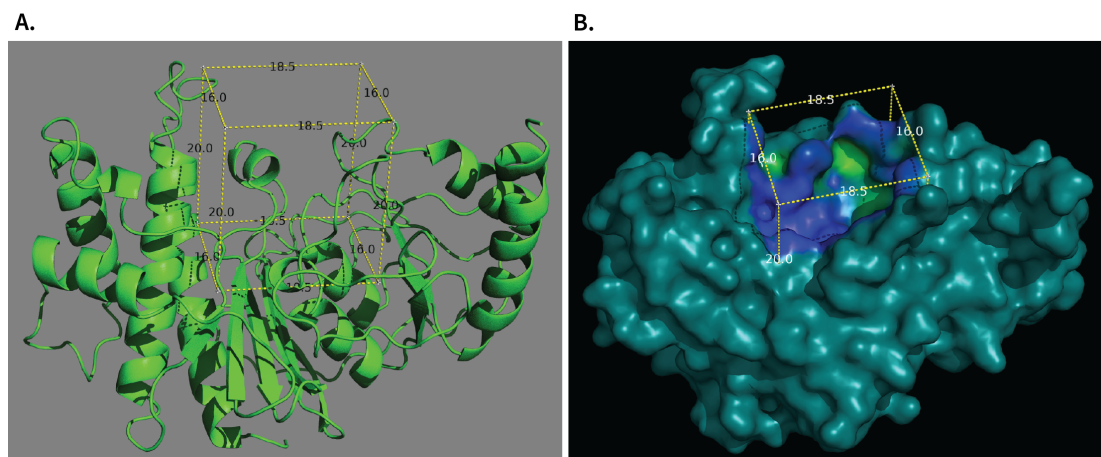


Figure 5.2: Active site encompassed cropped volume shown on AtSAL1, active site volume shown by yellow-dashed box, dimensions are in angstroms. (A) Cartoon representation. (B) Surface representation of AtSAL with surface charge shown for the active site-volume only. Surface charge was calculated at pH 7.6 with APBS at default values. Charge range was +5 to -5 kT/e. Blue indicates areas of negative charge. Green indicates areas of positive charge. Cyan (outside of the bounding box) is not displaying charge.

Coordinates (in angstroms) of the top plane were determined using the following code as the active site prism was not aligned with the global coordinates (so simple addition of distance to the bottom plane would skew the prism). Details on the coordinate calculation and final coordinates are provided in Appendix B Active site coordinates B.2.

The grid metadata (origin coordinates, delta spacing vectors, and grid dimensions) was parsed from the file header for each electrostatic potential map in OpenDX format. The full three-dimensional scalar field was then extracted and reshaped according to the specified dimensions. To determine which grid points fell within the defined prism, a Delaunay triangulation of the eight prism corners was computed using SciPy's spatial module. To minimise computational cost, a bounding box approach was employed prior to point-wise testing. The minimum and maximum extents of the prism were calculated in world coordinates, then transformed to grid index space using the origin and delta spacing. These indices were clipped to the valid grid range and used to define a sub-region containing all potentially relevant points. Only grid points within this bounding box were tested for prism membership. A new grid was initialised with dimensions matching the bounding box. For each position in this new grid, the corresponding world coordinates were calculated by transforming the grid indices using the new origin and delta spacing. If the Delaunay triangulation indicated the point lay within the prism (simplex index ≥ 0), the electrostatic potential value from the original grid was copied to the corresponding position in the cropped grid. Points outside the prism were assigned zero values. The cropped data was written to a new OpenDX format file, preserving the original delta spacing but with an updated origin corresponding to the bounding box minimum and reduced grid dimensions. This process was parallelised across all input files using Python's ProcessPoolExecutor with separate Delaunay triangulations computed independently for each worker process.

Maximum Intensity Projection

Three-dimensional frequency-filtered electrostatic fields were reduced to two-dimensional representations using **Maximum Intensity Projection (MIP)** along the z-axis. For each filtered field $\phi_{\text{filtered}}(x, y, z)$ at each frequency band, the projection was computed as $Z(x, y) = \max_z |\phi_{\text{filtered}}(x, y, z)|$, where the maximum was taken over all z-axis values at each (x, y) position. This approach preserved the strongest electrostatic features regardless of their depth within the protein structure. The resulting two-dimensional fields were subsequently down sampled from the computational grid (128×128 points) to a visualisation grid using bi-cubic spline interpolation. All projections were normalised to a global maximum across both enzymes and all frequency bands to enable direct quantitative comparison. The **MIP** was applied independently to each of the three frequency bands (>10 Å, $5\text{--}10$ Å, and $2.9\text{--}5.0$ Å, generating two-

dimensional electrostatic landscapes for comparative analysis.

5.2.5 Analysis methods

Distance matrix construction

Pairwise distance measurements between protein electrostatic fields were organised into symmetric distance matrices for subsequent analysis. Matrices were constructed for each of six frequency bands from electrostatic fields computed on a 64^3 grid. Each matrix element D_{ij} represents the optimal transport distance between the electrostatic fields of proteins i and j at a given frequency band and pH value.

PERMANOVA

Permutational multivariate analysis of variance (PERMANOVA) was conducted following (Anderson, 2001) using a custom Python implementation. The analysis assessed whether categorical grouping variables (clade, taxa, subcellular localisation) explained significant variance in electrostatic distance matrices. Analysis was performed with 99,999 permutations using a fixed random seed for reproducibility. Any groups containing fewer than three samples were excluded. The coefficient of determination (R^2) was calculated as the ratio of between-group to total sum of squares. Multiple testing correction was applied using the Benjamini-Hochberg false discovery rate method (Benjamini and Hochberg, 1995) via `statsmodels.stats.multitest`. This tests that if group labels were randomly assigned, how often would we see separation this good or better.

Band correlation analysis

To assess the independence of structural information captured across spatial scales, pairwise Spearman rank correlations were calculated between the upper triangular elements of distance matrices for all frequency band pairs. High inter-band correlations indicate redundant structural information, whereas low correlations suggest that different spatial scales capture independent aspects of electrostatic variation.

Hierarchical clustering

Hierarchical clustering was performed independently for each frequency band using average linkage (UPGMA) via `scipy.cluster.hierarchy`. Dendrograms

were cut to produce $k = 3$ clusters for comparison with biological groupings. Ward's linkage was additionally employed for robustness assessment.

The quality of clustering was assessed a ratio metric:

Between-to-within cluster distance ratio: For a given clustering $\mathbf{c}$, within-cluster and between-cluster distance sets were defined as:

$$\mathcal{D}_{\text{within}} = \{D_{ij} : c_i = c_j, i < j\} \quad (5.1)$$

$$\mathcal{D}_{\text{between}} = \{D_{ij} : c_i \neq c_j, i < j\} \quad (5.2)$$

The clustering quality metric Q was calculated as:

$$Q = \frac{\bar{d}_{\text{between}}}{\bar{d}_{\text{within}}} = \frac{\frac{1}{|\mathcal{D}_{\text{between}}|} \sum_{d \in \mathcal{D}_{\text{between}}} d}{\frac{1}{|\mathcal{D}_{\text{within}}|} \sum_{d \in \mathcal{D}_{\text{within}}} d} \quad (5.3)$$

where values $Q > 1$ indicate well-separated clusters.

Statistical validation through permutation testing

Test Statistic Definition

The association between clustering patterns and biological annotations was quantified using annotation homogeneity within clusters. For categorical annotations, the test statistic T was defined as:

$$T = \sum_{j=1}^k \sum_{a \in \mathcal{A}} p_{ja}^2 \quad (5.4)$$

where p_{ja} represents the proportion of proteins in cluster j with annotation value a , and $\mathcal{A}$ is the set of unique annotation values.

For continuous annotations (pH values), the test statistic was the between-cluster sum of squares:

$$T_{\text{continuous}} = \sum_{j=1}^k n_j (\bar{x}_j - \bar{x})^2 \quad (5.5)$$

where n_j is the size of cluster j , $\bar{x}_j$ is the mean annotation value in cluster j , and $\bar{x}$ is the overall mean.

Null Distribution Generation

Statistical significance was assessed through permutation testing with $N = 1000$ permutations. For each permutation π :

$$T^{(\pi)} = T(\mathbf{c}, \pi(\mathbf{a})) \quad (5.6)$$

where $\pi(\mathbf{a})$ represents a random permutation of the annotation vector while keeping cluster assignments fixed.

The empirical p-value was calculated as:

$$p = \frac{1 + \sum_{i=1}^N \mathbb{I}(T^{(\pi_i)} \geq T_{\text{observed}})}{N + 1} \quad (5.7)$$

where $\mathbb{I}(\cdot)$ is the indicator function.

Effect size was quantified using Cohen's d :

$$d = \frac{T_{\text{observed}} - \mu_{\text{null}}}{\sigma_{\text{null}}} \quad (5.8)$$

where μ_{null} and σ_{null} are the mean and standard deviation of the null distribution.

Controls

Paralogue distance comparison

To validate that the Fourier decomposition captures genuine evolutionary signal rather than methodological artefacts, pairwise electrostatic field distances were classified as within-paralogue (both maps are from the same paralogue) or between-paralogue (maps from different paralogues) based on known evolutionary assignments. For each frequency band, the test statistic was calculated as the difference between mean between-paralogue distance and mean within-paralogue distance. Positive values indicate greater similarity among evolutionarily related proteins. A null distribution was generated by randomly permuting paralogue labels across the 2,350 electrostatic maps (506 proteins $\times$ 5 pH conditions) for 1,000 iterations, preserving the number of maps per paralogue while destroying the relationship between evolutionary identity and electrostatic similarity. One-tailed p-values were calculated as the proportion of null differences equal to or exceeding the observed difference. Effect sizes were quantified using Cohen's d , calculated as the standardised difference between the observed statistic and the null distribution mean. Analysis was performed across six frequency bands; as p-values of all bands exceeded the $\alpha = 0.001$ threshold (multiple comparison correction: all bands ($n=6$) exceeded $\alpha = 0.006$).

Progressive noise robustness

Signal robustness was evaluated by introducing bounded multiplicative noise at levels $\sigma \in \{0.00, 0.02, 0.05, 0.10, 0.15, 0.20, 0.30, 0.40, 0.50, 0.75, 1.00\}$. The per-

turbed distance matrix was generated using mean-corrected log-normal noise:

$$\tilde{D}_{ij} = D_{ij} \cdot \exp(\tilde{\epsilon}_{ij}) \quad (5.9)$$

where $\tilde{\epsilon}_{ij}$ was drawn from a truncated normal distribution with mean correction to ensure $\mathbb{E}[\exp(\tilde{\epsilon}_{ij})] = 1$:

$$\epsilon_{ij} \sim \mathcal{N}\left(-\frac{\sigma^2}{2}, \sigma^2\right), \quad \tilde{\epsilon}_{ij} = \text{clip}\left(\epsilon_{ij}, -\frac{\sigma^2}{2} - 3\sigma, -\frac{\sigma^2}{2} + 3\sigma\right) \quad (5.10)$$

where ϵ_{ij} is the raw noise sample drawn from a normal distribution with mean $-\sigma^2/2$ (providing mean correction) and variance σ^2 , and $\tilde{\epsilon}_{ij}$ is the truncated noise obtained by clipping values to lie within $\pm 3\sigma$ of the distribution mean. The mean correction ensures $\mathbb{E}[\exp(\tilde{\epsilon}_{ij})] \approx 1$, so that noise application preserves expected distance magnitudes on average.

5.2.6 Statistical Significance Assessment

All p-values were set with significance thresholds at $\alpha = 0.05$ for standard significance, $\alpha = 0.01$ for high significance, and $\alpha = 0.001$ for very high significance. False discovery rate correction was used.

5.2.7 Similarity networks

Sequence similarity networks (as distance matrices) were generated as per Chapters 3 and 4. The structural similarity network was created using FOLDSEEK version 10 (van Kempen et al., 2024). ‘Easy search’ mode was used and run locally with the code provided in Appendix B Structural similarity network construction B.3.

All networks (distance matrices) were filtered to the 0.08 quantile by the respective distance metric unless otherwise stipulated. The distance matrices were then imported into Cytoscape 3.10.4 for formatting and visualisation. Clustering was performed with the Perfuse Force Directed layout method.

5.3 Results

5.3.1 Approach 1: Comparisons of 2-dimensional image active site surface charge

Analytical development and computational pipeline

Following the logic previously outlined where primary enzyme function derives directly from the electrostatic field, only indirectly from structure and at a further level of abstraction from sequence, I developed a novel approach to measure differences between nucleotide phosphatases for classification purposes. I decided to compare electrostatic maps as a proxy for function. Since electrostatic maps share the same structural properties as images, I chose to use [Earth Mover's Distance \(EMD\)](#), an established method for comparing images. The computational complexity of three-dimensional comparisons warranted a proof-of-concept. I therefore created a simple pipeline to compare two-dimensional images of the active site surface charge.

This test was limited to the active site as this region performs catalysis (the enzyme's primary function) and should prove most informative for functional comparisons. Comparing the entire enzyme surface would have required map projection, the same concept used for flattening a globe to a flat map. Since spheres cannot flatten without distortion, a projection method becomes necessary to minimise distortion. This adds unnecessary complexity for this proof-of-concept. While a simplified approach, these comparisons would demonstrate the approach had merit if structure was detectable in the data, while limited by likely lacking the accuracy needed to address specific biological questions. The absence of structure cannot prove the negative. Absence of data-structure might result from significant noise in comparing two-dimensional active site images. This limitation means it can not show a false negative result. A positive result, however, would provide quick validation of the concept and create a foundation for further development.

STRUCTURE GENERATION

From the set of 543 sequences (including the outgroup), [AF2](#) successfully predicted 506 structures, while 37 failed. I validated [AF2](#)'s performance against the experimentally characterised AtSAL1 crystal structure (PDB:8F9Y) by aligning in PyMol and visually comparing them as seen in [Figure 5.3](#) which showed minimal differences outside of loop regions. Of note, while they do not perfectly match, crystal structures also contain error from determination described as resolution (8F9Y = 2.60 Å resolution), and that crystal structures are determined from precipitated protein that is no longer free to move within its solvent which may not represent its natural confirmation, nor does a static representation show *in vivo* conformational dynamics. As such, the minimal differences

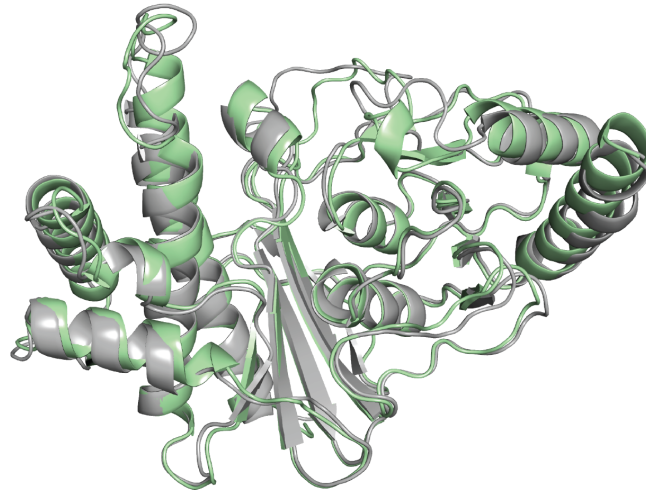


Figure 5.3: AtSAL1 crystal structure–AlphaFold2 structural model comparison. Structure are represented in cartoon format, generated in PyMol. Grey denotes the experimentally determined crystal structure (PDB:8F9Y). Green is the AlphaFold2 structural model (pLDDT of 91.6) created using Colabfold running the the AlphaFold2 model, using default parameters except for no template used (*de novo* prediction) and the structure was relaxed using AMBER force field. The sequence had the experimentally determined transit peptide removed prior to modelling.

between the structures were assessed to be within experimental bounds given all structures were *de novo* modelled regardless of whether an experimentally determined structure was available to ensure consistency within the dataset. I also manually examined a random selection of 10% of the generated structures aligning them in PyMol to the AtSAL1 crystal structure (not included) to ensure there were no obvious signs of misfolding that had passed the pLDDT filter where sequences with pLDDT < 0.7 were removed (no sequences were under this threshold). The distribution of pLDDT scores of each paralogue are seen in Figure 5.4 with the means of AHL and CNP being ‘excellent’ and SAL ‘very good’ as defined by (Jumper et al., 2021; Tunyasuvunakool et al., 2021). SAL would have been expected to have the highest mean pLDDT score as it was first discovered through sequence homology with ScHAL2 suggesting it is more closely related to better studied orthologues and should therefore be better represented in the training data.

ELECTROSTATIC CALCULATION

Electrostatic maps were calculated for all 506 structures at five pH values (7.0, 7.3, 7.6, 7.9, 8.2) using the Poisson-Boltzmann approach implemented through PDB2PQR and APBS.

Using a custom Python script to control PyMOL (see Methods), I generated images of active sites from identical perspectives for each homologue in the pre-aligned nucleotide phosphatase dataset. PyMOL’s inbuilt surface charge representation creates these images by colouring by charge according to where

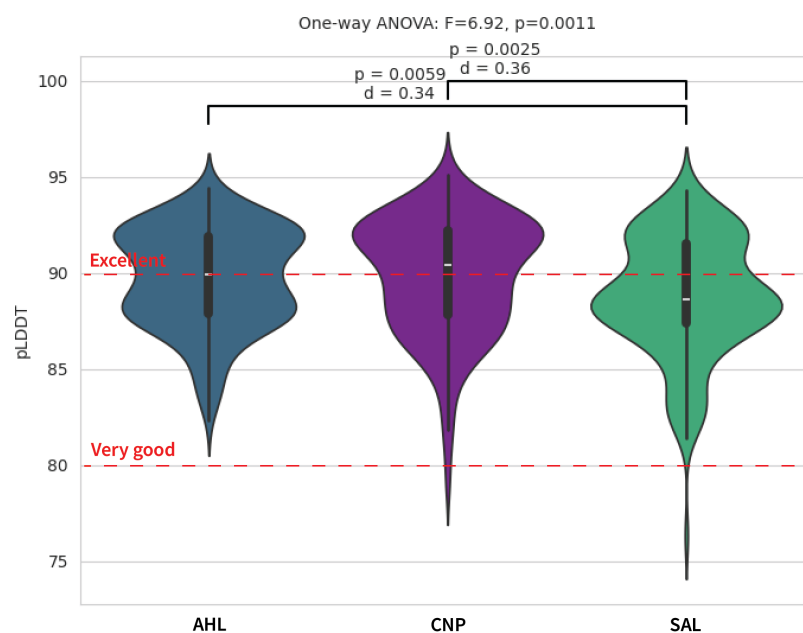


Figure 5.4: AlphaFold2 predicted structures pLDDT scores shown by paralogue as a violin plot. Left plot is AHL, middle is CNP, right is SAL. Statistical comparison used Tukey HSD with p-values adjusted by false discovery rate, effect sizes are Cohen's d . $N = \text{AHL:187, CNP:158, SAL:161}$. No significant difference between AHL and CNP, highly significant difference between SAL and AHL, and CNP with effect sizes of 0.34 and 0.36 respectively.

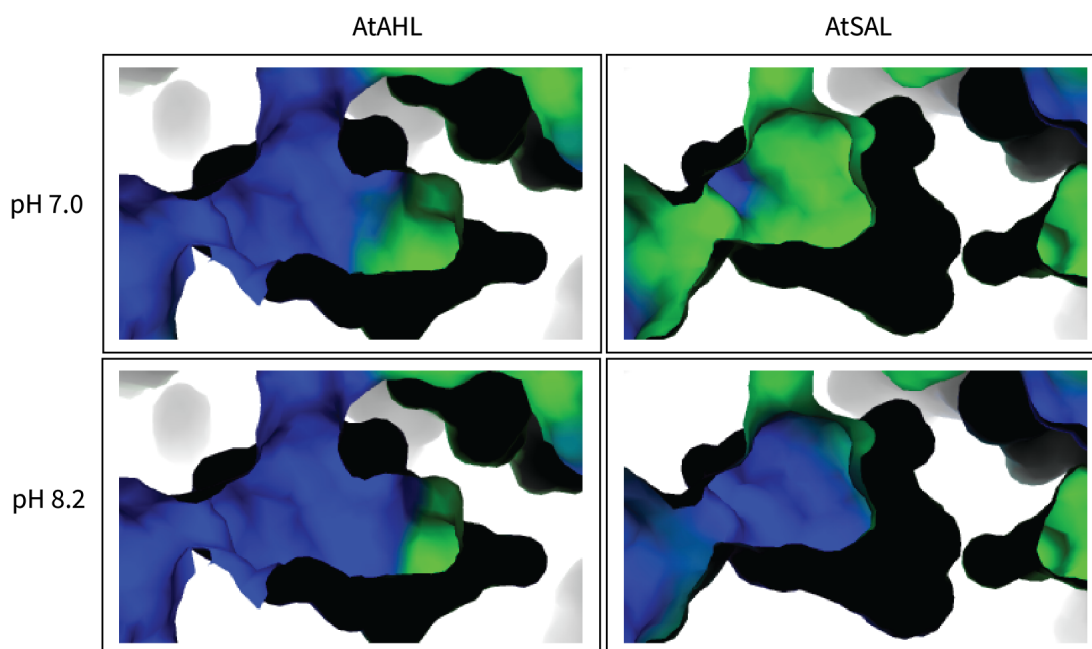


Figure 5.5: Example approach 1 active site surface charge images showing AtAHL (left) and AtSAL1 (right) at pH 7.0 (top) and 8.2 (bottom). Charge range was +5 to -5 kT/e. Blue indicates areas of negative charge. Green indicates areas of positive charge. Black represented neutral charge or obstruction. White shows voids. Images were generated in PyMol with the following settings: max_ups=0, use_shaders=0, sphere_mode=9, specular=0.0, ray_shadows=0, images generated with ray tracing.

the electrostatic map intersects the original structure's surface. A diverging colour map was used where one colour's intensity represented negative charge

strength, while the other colour indicated positive charge. Black represented neutral charge or obstruction. This scheme ensured colour remained directly and proportionately related to charge. Common multi-colour maps used in scientific figures, such as ‘viridis’, would have introduced artefacts with this simple approach. The script then captured images using PyMOL’s inbuilt functionality. Programmatically generating these images from identical perspectives created rectangular images of equal size that allowed direct comparison. Example images showing AtAHL and AtSAL1 at pH 7.0 and 8.2 are presented in Figure 5.5 showing the variation in both active site structure and charge between enzymes at varying pH. This figure captures the difficulties in determining ‘similarity’ between the active sites given charge and structural differences and the loss of dimensionality. PyMol was chosen over other visualisation programs such as VMD (Humphrey et al., 1996) due to my familiarity with the tool and the ease of programmatically controlling it using Python, the programming language I had been learning throughout my PhD; this enabled the programmatic alignment of all 506 structure, the generation of the surface charge representation of the active sites, and then image capture.

DISTANCE CALCULATION

I considered a novel approach of measuring differences between these active site surface charge images using **EMD**. **EMD** represents a fundamentally different paradigm for comparing three-dimensional electrostatic maps. Originally developed for image comparison and optimal transport problems¹ (Rubner et al., 2000), **EMD** conceptualises each electrostatic map as a distribution of “mass” where positive potentials represent sources and negative potentials represent sinks. Rather than requiring point-to-point correspondence like **RMSD** or correlation indices, **EMD** calculates the minimum “work” required to transform one distribution into another (see Figure 5.6) where efficiency means minimising the total distance that mass must be transported. Mathematically, this becomes an optimisation problem solved through linear programming. **EMD** can be calculated for discrete or continuous distributions. This approach uses continuous distributions and that equation is defined as

$$\text{EMD}(\mu, \nu) = \inf_{\gamma \in \Pi(\mu, \nu)} \int_{X \times Y} c(x, y) d\gamma(x, y) \quad (5.11)$$

where:

¹An optimal transport problem is a mathematical framework for finding the most efficient way to transform one probability distribution into another

- μ and ν are probability measures on spaces X and Y respectively
- $\Pi(\mu, \nu)$ is the set of all joint probability measures (transport plans) with marginals μ and ν
- $c(x, y)$ is the cost function defining the cost to transport mass from x to y
- γ is a transport plan that satisfies the marginal constraints

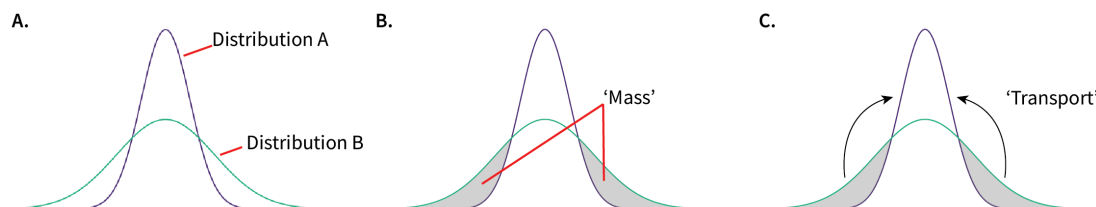


Figure 5.6: Earth Mover's Distance (EMD) quantifies the minimum work required to transform one distribution into another. (A) Two electrostatic distributions with different shapes and peak positions. Distribution A (purple) shows a tall, narrow peak while Distribution B (green) displays a shorter, broader profile. (B) The mass difference between distributions, shown as grey shading, represents the amount of electrostatic potential that must be redistributed. (C) The optimal transport plan minimises the total distance that mass travels, with curved arrows indicating the direction and magnitude of mass movement required to transform Distribution A into Distribution B. In protein electrostatic comparisons, this approach captures both the magnitude of differences and their spatial relationships without requiring point-to-point correspondence.

The algorithm calculates transport plans that specify how much “charge” moves from each position in the first protein to each position in the second protein, with the constraint that all charge must be accounted for and the total transport cost must be minimised. This formulation naturally handles structural variation that confounds other methods. When a charged residue shifts position by 2 Å between homologous structures, **EMD** accounts for this displacement through a small transport cost rather than treating it as a complete mismatch. An inserted loop carrying additional charged residues increases the transport cost proportionally to both the amount of charge and the distance to the nearest matching features. The method provides a continuous measure of similarity that degrades gracefully with increasing structural divergence rather than catastrophically failing when alignment becomes impossible. The mathematical foundation of **EMD** in measure theory provides rigorous theoretical guarantees about the metric properties of the resulting distance matrix (Rubner et al., 2000). Unlike correlation coefficients that can produce non-intuitive results when comparing multiple proteins, **EMD** satisfies the triangle inequality. This means that the distance from protein A to protein C cannot exceed the sum of distances from A to B and B to C, a property essential for meaningful clustering and phylogenetic analysis. The method also preserves absolute magnitude information that normalised indices obscure. A protein with a strongly charged active site will show appropriate distance from a weakly charged variant, whereas correlation indices might show high similarity if the

spatial patterns match.

The **EMD** algorithm then calculated pairwise distances between AtSAL1 at all pHs and the rest of the dataset. The logic at the time was to identify homologues that were functionally similar to AtSAL1. While computationally faster, this introduced additional complexity and made comparisons between other homologues not possible unless couched in terms of their relationship to AtSAL1. This itself prevents triangle equality, and partially undermines the approach. I coded the pairwise comparisons to be done non-redundantly. A simpler implementation would have generated comparisons between enzyme A and B as well as B and A, which are equivalent for my purposes. This redundancy would have created significant extra computational cost that scaled poorly, while also requiring post-computation filtering of redundant comparisons. The pairwise comparisons created a distance matrix using the same approach as **SSNs** in Chapters 3 and 4. The **SSNs** relied on BLAST similarity scores to provide 'distances' between sequences. Similarly, **EMD** provided distance scores. I then filtered and processed this network before visualising it in Cytoscape using the same methods as the **SSNs** in earlier chapters, creating an 'image similarity network'. Taxa served as features for node colouring. If network nodes clustered together qualitatively, this would suggest data structure exists and the **EMD** method can identify similar electrostatic maps. Figure 5.7 shows this clustering, where common features (same colour) cluster in a non-random pattern. As mentioned above the comparisons between homologues are not actually directly relatable and are only comparable in their relationship to AtSAL1, however at the time I had not fully understood this. Regardless, I determined qualitative comparison sufficient for a proof-of-concept given the visible level of clustering observed, without using detailed statistical comparisons to confirm non-random clustering.

Biological assessment

While this pilot test showed promise, surface charge representations present two core limitations. First, we observe the active site from only one perspective. This resembles looking into a cave where twists obscure sight lines, providing only a partial view of the cave's contents. Active sites in this dataset vary structurally, with some more obstructed than others, leading to differential error levels between active sites. This information loss demonstrates the value of a three-dimensional comparison approach (Figure 5.5). Second, as discussed earlier in the chapter, surface charges represent only part of an active site's functional landscape despite their critical importance for understanding cata-

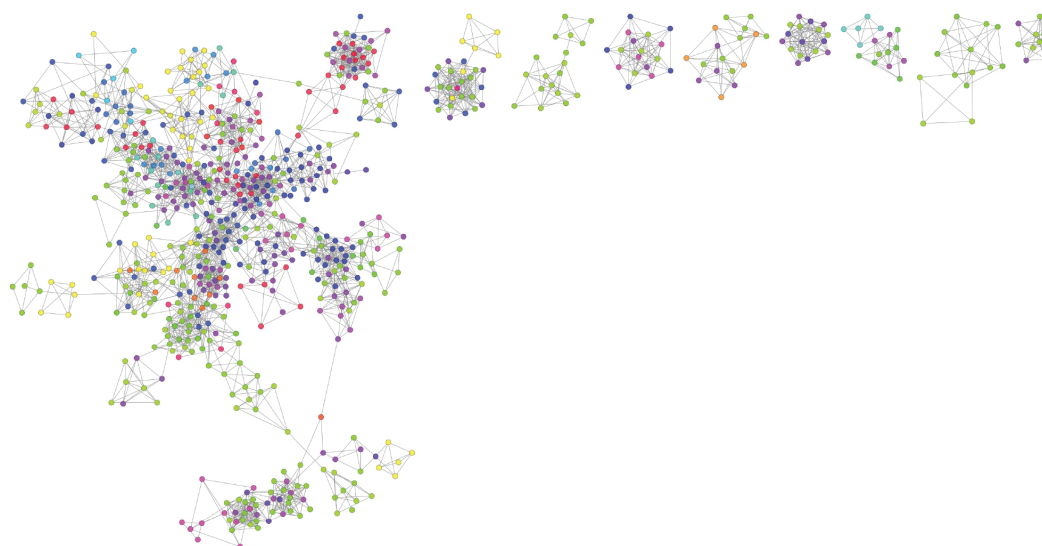


Figure 5.7: Active site image similarity network showing structure in the data (coloured nodes distributed non-randomly) where nodes (images) are coloured by taxa.

lytic mechanisms. Non-surface charges remain important for driving substrate channelling into the binding pocket and carrying global electrostatic features. Additionally, the overall field within the active site results from the protein's complete structure-charge layout. Non-surface components of the active site electrostatic field therefore contain allosteric information about the protein as a whole. Surface or near-surface representations lose much of this information.

The apparent partial success of this proof-of-concept led me to integrate comparisons of full active site electrostatic maps using the [EMD](#) method in my second approach.

5.3.2 Approach 2: Comparisons of three-dimensional active site electrostatic maps

Analytical development and computational pipeline

Following electrostatic calculation, all structures required alignment to the AtSAL1 reference structure to enable consistent active site extraction. Initial attempts using PyMol's built-in alignment functions took 3-6 minutes per structure when performed manually, necessitating another automation script (see Methods). This approach failed and required troubleshooting where I determined that the electrostatic map files couldn't be directly aligned post-calculation. The grid-based nature of [APBS](#) output files means that spatial transformations would require interpolation between grid points, introducing numerical errors. Instead, I modified the workflow to first align all protein structures to the reference, then generate electrostatic maps from these pre-aligned coordinates. Act-

ive site extraction presented unexpected geometric challenges. The active site region wasn't aligned with the global coordinate axes. Consequently, the rectangular bounding box needed for cropping ran diagonally through the three-dimensional electrostatic grid. This misalignment created ambiguity at boundaries where any grid points partially inside the cropping region required decisions about inclusion or exclusion. This resulted in the cropped maps having varying dimensions, preventing direct numerical comparison of electrostatic properties. The challenge was comparing electrostatic fields across different protein structures when each structure had been calculated on a different grid with different spacing and orientation. Direct comparison was not possible because the grid points did not correspond between structures. The solution came from Delaunay triangulation (Cheng et al., 2016), a computational geometry technique commonly employed in computer graphics for terrain rendering and in computational fluid dynamics simulations. The method constructs a mesh of non-overlapping tetrahedra from a set of discrete points, enabling interpolation of values at any location based on the surrounding measured points. Delaunay triangulation operates by connecting points to form tetrahedra according to a specific geometric criterion. For each tetrahedron in the mesh, a circumsphere can be defined (a sphere passing through all four vertices). The Delaunay criterion ensures that no other points from the original dataset lie within the interior of any circumsphere. This property generates a well-balanced mesh that yields smooth, numerically stable interpolation with minimal artefacts. By applying Delaunay triangulation to the electrostatic grid points from each structure, I resampled the electrostatic field onto a standardised coordinate system aligned with the active site. This approach enabled quantitative comparison across all electrostatic maps and enabled recompiling of the cropped maps with the standard, required, grid layout.

The shift from two-dimensional to three-dimensional EMD comparisons proved technically straightforward to implement. Unfortunately, it became computationally intractable.

The EMD computational complexity typically scales as $O(N^3)$ where N represents the number of points:

$$\text{2D EMD complexity: } O((n^2)^3) = O(n^6) \quad (5.12)$$

$$\text{3D EMD complexity: } O((n^3)^3) = O(n^9) \quad (5.13)$$

$$\text{Increase factor: } n^9 / n^6 = n^3$$

This scaling means that comparing a 100×100 pixel two-dimensional image

to a $100 \times 100 \times 100$ voxel three-dimensional image would require roughly 1×10^6 times more computations.

Firstly I attempted to address this by processing the electrostatic maps through transformation into one-dimensional strings of charge values - referred to as approach 2.A. This significantly reduced the computational burden of calculating these distances to a trivial level, while also still only making pairwise comparisons to AtSAL1. I mapped these distances to the original phylogeny, which appeared to show some patterns (Figure 5.8) as did the similarity network (Figure 5.11). Unfortunately, both approaches contained fundamental flaws discussed below.

This necessitated a conceptual rethink which is referred to as approach 2.B. I re-worked the pipeline to make pairwise all-versus-all comparisons rather than just pairwise comparisons with AtSAL1. I also developed the approach to work with true three-dimensional electrostatic map comparisons instead of the one-dimensional transformed maps. This added several orders of magnitude increase in computational time to the point it would not be experimentally practical. This processing burden was addressed through more sophisticated computational methods including a less computationally expensive algorithm than EMD. The major improvement came through implementing the Sinkhorn algorithm (Cuturi, 2013). The Sinkhorn algorithm specifically addresses the EMD algorithm's limited scalability problem. This approach carried the costs of reduced precision (increased error) and coding complexity. I assessed the error cost as manageable since I could still achieve precision commensurate with the structural models' accuracy. A perfect solution remains not only computationally intractable but provides false precision beyond what the underlying structural model affords.

I replaced the EMD approach with the Sinkhorn implementation and performed all-versus-all comparisons for all 2,530 active site electrostatic maps. Each cropped active site electrostatic map contains 149,688 points from the three-dimensional map represented as voxels. Each point carries a numerical measure of electrostatic charge at that coordinate. This point count created too high a computational burden, so I approximated the charge distribution using a lower point count. This was achieved by sampling n points with the highest importance (importance sampling), where n remains configurable. I implemented importance sampling where points are selected with probability proportional to their distribution values. Points with higher probability mass (more 'charge' in the electrostatic field context) have greater likelihood of retention. The Sinkhorn algorithm requires non-negative probability distributions as input measures. Since electrostatic potential values can be negative, I ap-

plied a two-step normalisation procedure to transform them into valid probability distributions. First, shifting all values to ensure positivity by subtracting the minimum value and adding a small epsilon term for numerical stability. Second, normalising the shifted values by dividing by their sum, creating a probability measure that sums to 1. This transformation preserves the relative magnitudes of the original potential values while satisfying the mathematical requirements for optimal transport computation.

$$\tilde{\phi}_i = \phi_i - \min_j(\phi_j) + \epsilon \quad (5.14)$$

$$\mu_i = \frac{\tilde{\phi}_i}{\sum_{j=1}^N \tilde{\phi}_j} \quad (5.15)$$

where $\{\phi_i\}_{i=1}^N$ are the original potential values, $\epsilon = 10^{-10}$ prevents numerical instability, and $\mu = (\mu_1, \dots, \mu_N)$ is the resulting probability distribution with $\mu_i \geq 0$ and $\sum_i \mu_i = 1$.

This approximation approach aims to preserve the probability distribution's important features, unlike random sampling. In theory, points with negligible probability mass can be safely discarded, while the subsample better represents regions with high probability. In practice, however, this approach creates risk of artefacts where important regions such as the non-surface binding pocket region have lower probability mass but nevertheless contain important information. This approach to managing the computational burden directly influences the experimental outcome by intrinsically biasing the data. The aim is to ensure that the computational burden is manageable in an experimental context while minimising biases in the data. Whether importance sampling would affect the use-case through bias in the sampling remained unclear at that point.

Further processing time improvements were achieved through several different computational approaches to increase time-efficiency. While this may not seem relevant to the biological question of classifying enzymes, it proved complex, time-consuming, and absolutely critical. Initial CPU benchmarking of the three-dimensional Sinkhorn implementation revealed each calculation took approximately 10 s. This timing includes the complete workflow of loading data, processing, and writing results to file. This benchmarking was done by comparing two maps containing 7,500 points each, using $\epsilon = 10^{-4}$ with a minimum of 20 iterations and a maximum of 500 iterations. All calculations ran sequentially on the CPU. With approximately 3.2 million comparisons, a single run would take approximately 370 days. By running calculations in parallel on the GPU and using methods such as batching, I reduced this time to several

hours. As this pipeline was ultimately not used for analysis the full implementation details have not been included for brevity's sake, but the code is available via GitHub (see Methods). With run times at a manageable level, I could run the analysis for the entire dataset and generate the distance matrix.

Biological assessment

Approach 2.A

The one-dimensional comparison (approach 2.A) will have created artefacts from dimensionality reduction and distorted the active site's spatial data. This contradicted the core rationale for comparing three-dimensional active site electrostatic maps. This approach also suffered from the same limitation of measuring distance between AtSAL1 and all other homologues, rather than all-versus-all comparisons. The five electrostatic maps across the pH 7.0-8.2 range for each homologue were not directly comparable to each other as the distances all related back to AtSAL1 which created further complications in analysis.

This resulted in less informative results that proved more difficult to interpret (Figure 5.8). Initial statistical analysis showed no significant difference in predictive power between clustering by paralogue, pH, taxa, or localisation compared to randomised data. Table 5.1 shows the paralogue agreement/clustering results with no high significance. Failing this metric is sufficient to fail the approach as this is the highest level of classification available in the dataset and for which I have sequence-homology – therefore putative – benchmarking. As such, for brevity I have not included the less robust measures such as clustering by taxa or pH environment.

Table 5.1: Approach 2 paralogue agreement results

Paralogue	n	Actual Agreement	Random Agreement	P-value
Algal AHL	36	0.495	0.605 ± 0.095	0.8530
Basal AHL	65	0.362	0.398 ± 0.066	0.5760
AHL- α	43	0.364	0.325 ± 0.030	0.1750
AHL- β	43	0.380	0.415 ± 0.034	0.8470
CNP	158	0.546	0.506 ± 0.032	0.1160
SAL	160	0.469	0.460 ± 0.020	0.3330

Approach 2.B did use all-versus-all pairwise comparisons with true three-dimensional electrostatic map comparisons. The distance matrix was visualised as a dendrogram (clustering) heatmap (Figure 5.9). I attempted to create a similarity network with these results for simple qualitative comparison between this approach, approach 1 and methods such as sequence-homology. Unfortunately the variance in the distance values was minute. The distances between electrostatic maps or sequences (sequence similarity or percentage

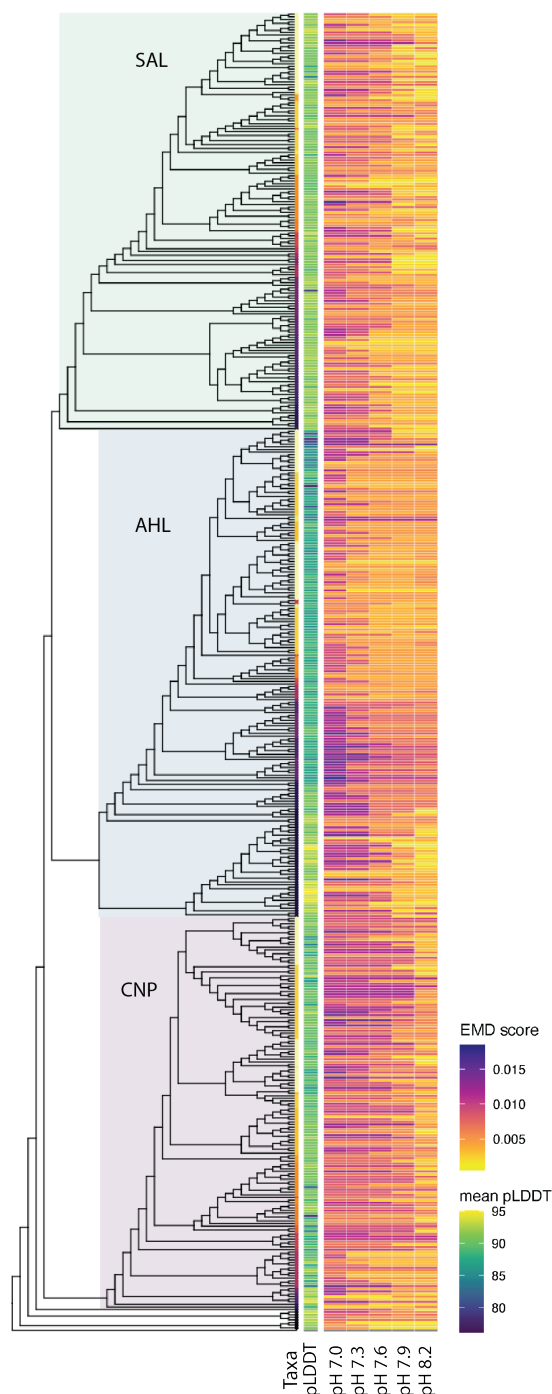


Figure 5.8: Heat map of pairwise comparisons with AtSAL1 across a pH range of 7.0–8.2, mapped against the phylogeny with branch lengths = 0, tips are coloured by Order. Structure pLDDT scores are shown between the phylogeny and the heatmap. Heatmap shows Earth Movers Distance where a lower score (yellow) indicates higher similarity.

identity) need to be filtered to the most informative for visualisation in a similarity network. When comparing electrostatic maps in a similarity network I filtered by the 0.08 quantile to find the smallest pairwise distances (closest relationships) between the electrostatic maps (Figure 5.10). Iterative testing showed this was the optimal for showing clustering of the paralogues across

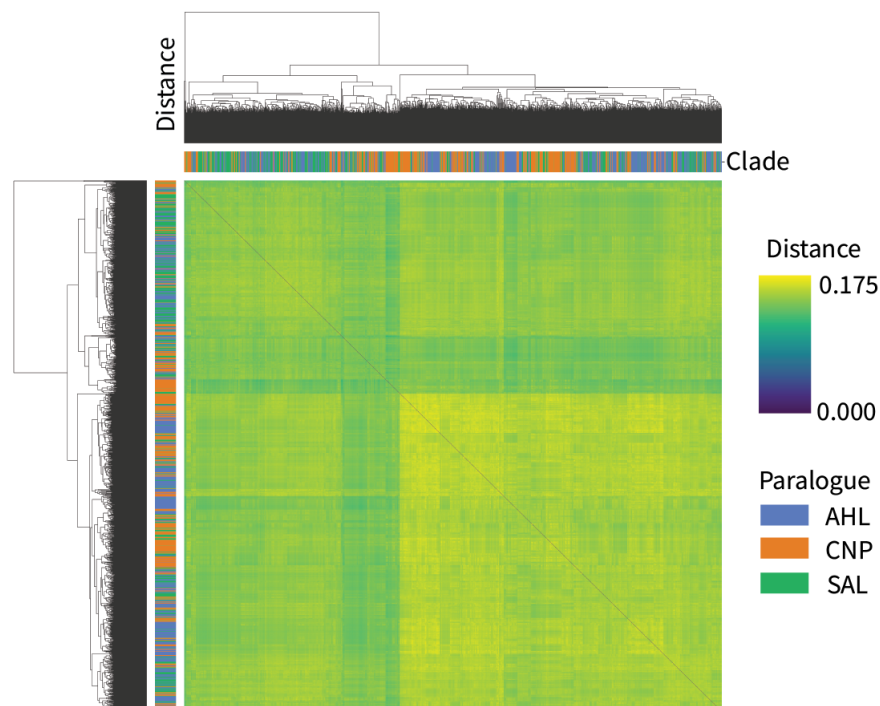


Figure 5.9: Active site electrostatic distance heatmap and dendrogram using the Sinkhorn algorithm using 7,500 points for subsampling, marginal constraint tolerance threshold 1×10^{-4} , minimum and maximum iterations for convergence 20/500, and normalised by calculating the maximum distance and divides all distances by this value then multiplies by the epsilon of 0.01. The dendrogram was mapped to the parologue labels showing degree of clustering by parologue. Dendrogram was clustered by hierarchical clustering (see Methods section).

the dataset. Unfortunately the variance in this particular network for the central 0.95 quantile was $\sim 6.73 \times 10^{-6}$ and far too small for Cytoscape's layout algorithms to separate the clusters (Figure 5.11). This is also evident in the heatmap (Figure 5.9) where we see little variation in colour across the heatmap.

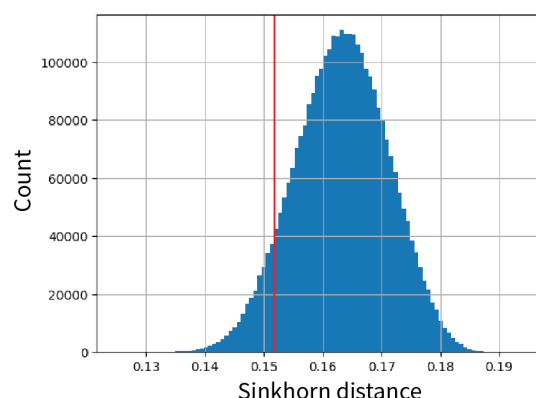


Figure 5.10: Approach 2 Sinkhorn all-versus-all pairwise distance distribution calculated by Sinkhorn optimal transport. Values to the left of the red line represent the 0.08 quantile of values used to generate the electrostatic map similarity networks.

While Figure 5.11 is not informative Figure 5.9 shows clustering which, while

weak, appears to be non-random. This is likely due to three principle causes. The first is simply noise in the data. The second cause could be the poor sampling of the non-surface regions caused by the potentially ironically named ‘importance sampling’ outlined above. This could lead this method to be *de facto* similar in approach to the PIPSA method near-surface method outlined earlier in this chapter. The third cause is likely because the Sinkhorn algorithm is taking the net charge distribution while fine changes in the charge distribution in the electrostatic map are being ‘averaged out’. This approach takes the importance sampled probability mass as a whole and calculates the mass transport cost across for the entire map thereby loosing spatial variation. The MatchTope method addresses this by ‘chunking’ the map into small pieces for comparison.

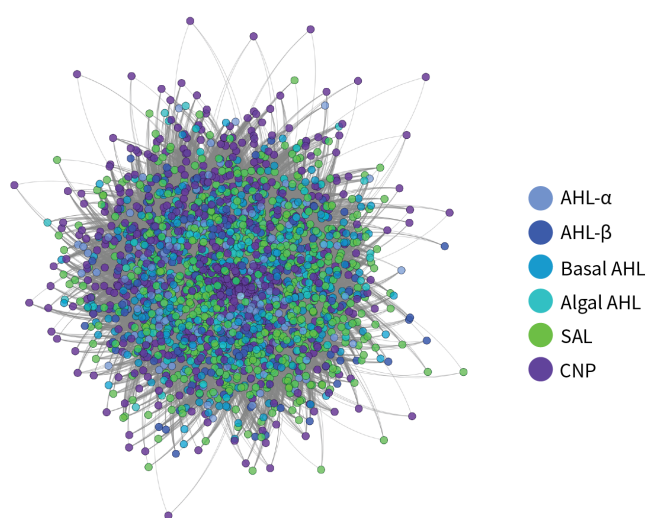


Figure 5.11: Approach 2 electrostatic map all-versus-all pairwise comparison similarity network using Sinkhorn distance with edges filtered to 0.08 quantile and coloured by paralogue.

All three of these problems are able to be solved through the novel application of a Fourier transformation which was applied for approach 3.

5.3.3 Approach 3: Fourier decomposition of three-dimensional active site electrostatic maps

Approach 2 compared the total charge contribution between all the enzymes, pairwise but showed no substantial ability to discriminate between paralogues. This suggests that different nucleotide phosphatases could share highly similar net charge yet catalyse different reactions with different efficiencies. Regions of positive charge in the phosphatases’ active sites will attract their negatively charged substrates which in the case of AtSAL1 and AtAHL this is PAP (Quintero et al., 1996; Chan and Warwicker, 2009; Gil-Mascarell et al., 1999)

². AtSAL1 degrades PAP efficiently unlike AtAHL (Quintero et al., 1996; Gil-Mascarell et al., 1999). This difference therefore must not arise from the total difference in charge which is supported by Approach 2's results. Approach 2 calculated optimal transport distances directly between electrostatic maps, producing a single distance metric that captured the overall difference in charge. The method was unable to discriminate between paralogues, suggesting that SAL, AHL, and CNP share similar net charge distributions despite their functional differences. If overall charge cannot distinguish between the paralogues, then functional differences must arise from how charge is distributed spatially and at different scales, if noise is not overriding all signal.

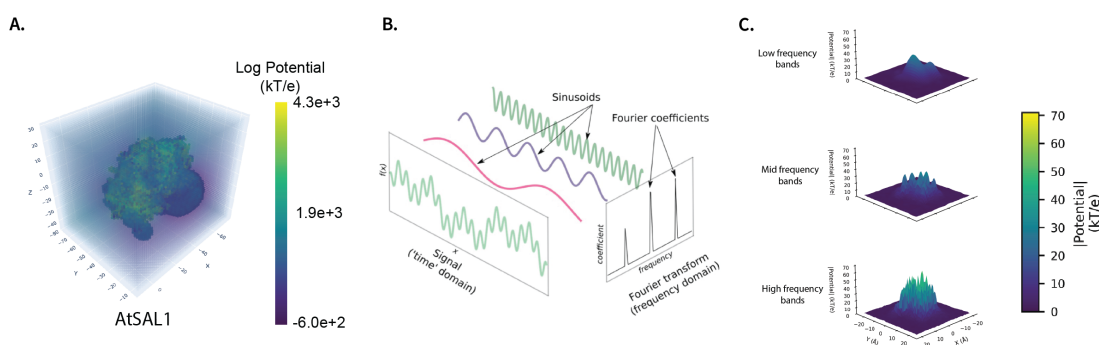


Figure 5.12: (A) Three-dimensional image of AtSAL1's electrostatic map at pH 7.9, voxels are colour by charge potential (log kT/e). (B) Fourier decomposition transforms complex spatial patterns into constituent frequency components. A complex signal in the spatial domain (left, green trace) decomposes into individual sinusoidal components of varying frequencies (centre, coloured waves). Each sinusoid corresponds to a specific spatial frequency with an associated amplitude coefficient (right, frequency spectrum). In protein electrostatic analysis, this transformation separates global charge distributions (low frequencies) from local active site features (high frequencies), enabling scale-specific comparison of electrostatic patterns. Adapted from Mahakaran et al., (2025). (C) Frequency decomposition of the electrostatic landscape of AtAHL at pH 8.2. Maximum intensity projections of frequency-filtered electrostatic potential fields are shown for three spatial frequency bands: low frequency ($\gg 10$ Å), mid frequency (5–10 Å), and high frequency (2.9–5.0 Å). The x- and y-axes represent spatial coordinates in angstroms relative to the active site centre, and the z-axis shows the maximum absolute potential magnitude in kT/e.

Fourier decomposition accounts for spatial arrangement of charge and breaks it down into different spatial scales which can be analysed separately. Different spatial scales associate with different features. At a fine scale (high-frequency) charges associated with individual residues or atoms can be resolved, while at larger scales (low-frequency) charges associated with the overall active site can be compared. A common use of Fourier decompositions is isolating frequencies in sound, such as adjust high, mid, and bass frequencies in music or removing noise from an audio recording. In a similar way to this Fourier decomposition can be used to isolate specific features in an electrostatic map

²While AHL can degrade PAP there is no experimental evidence that PAP is AHL's primary substrate.

(Figure 5.12A) by breaking complex spatial patterns into components at different wavelengths (Duoandikoetxea, 2024; Cooley et al., 1969) (Figure 5.12B). Applied to electrostatic maps, low frequencies capture the broad distribution of charge across the protein, mid frequencies capture the arrangement of charge within functional regions, and high frequencies capture the positioning of individual residues and hydrogen bonding networks. Isolating each frequency band enables analysing what differentiate the paralogues from each other, such as overall charge, charge distribution in active site geometry, or the fine positioning of catalytic residues.

Low-frequency components represent gradual electrostatic variations extending over tens of angstroms. These capture the gentle slopes and broad valleys of the electrostatic field, potentially encoding global properties such as the overall dipole moment or long-range gradients that guide substrates toward active sites. These global properties, missed by methods that focus on surface or near-surface charge, can be functionally critical. Superoxide dismutase exemplifies this principle, exploiting long range electrostatic gradients to achieve diffusion-limited catalysis by electrostatically steering substrates faster than random diffusion allows (Getzoff et al., 1992).

Spatial scales and biological features

The relationship between frequency and wavelength determines what biological features each band – spatial scale – captures. Frequency (f) measures periodicity over distance in $\text{\AA}^{-1}$, while wavelength (λ) the inverse in $\text{\AA}$:

$$\lambda = \frac{1}{f}$$

For example, filtering frequencies between $0.15 \text{\AA}^{-1}$ and $0.20 \text{\AA}^{-1}$, isolates spatial patterns repeating every $5.0 \text{\AA}$ to $6.7 \text{\AA}$. This wavelength range matches the pitch of an alpha helix at approximately $5.4 \text{\AA}$ per turn. The translation from frequency to wavelength allows direct interpretation of which structural features likely contribute to electrostatic differences between proteins.

Implementation overview

This approach processes electrostatic maps by applying Fourier decomposition to break them into frequency bands. The number of points for each band are then reduced using deterministic point sampling before applying the Sinkhorn

algorithm for comparison.

Fourier decomposition of electrostatic maps

Given that different frequency components capture different structural features, we can define custom frequency bands to target specific scales of interest. Increasing the number of bands allows identification of the most informative bands for a given question, but this comes at computational cost. More bands means more pairwise comparisons since one band requires approximately 3.2 million comparisons so six bands require 6×3.2 million. This drives the need to optimise the number of bands and their boundaries to best meet biological analytical needs balanced against computational cost.

Noise filtering

Frequency decomposition allows the removal of information from spatial scales most associated with noise from different sources of error. In particular modelling error in the 1–2 Å range, minor side-chain placement error and numerical noise from [APBS](#) calculations as well as electrostatic maps point transformations can be isolated in the high-frequency ranges. In this study frequencies corresponding to <2.2 Å were discarded to minimise the impact of noise. This creates a low-pass filter that removes atomic-level noise and structural artefacts. This filters out biological features that relate to precise atomic positioning and is likely most associated with high-specificity substrate recognition and precise position of atomic-level catalytic features. Prior to this a power threshold was set at 10^{-12} times the maximum power to remove numerical noise before frequency point extraction. Since the power spectrum is the square of [Fast Fourier Transform \(FFT\)](#) magnitudes computed in single-precision floating point (float32, machine epsilon $\approx 10^{-7}$), squaring yields a precision floor of approximately 10^{-14} . The threshold of 10^{-12} relative to maximum power therefore lies two orders of magnitude above this floor, providing a conservative buffer that excludes values arising from floating-point accumulation during [FFT](#) computation. This still retains weak but genuine spectral features.

Distance of frequency components

Separating electrostatic fields into frequency bands transforms comparison from a single net charge distribution metric to a multi-dimensional characterisation. Each band captures features at specific spatial scales with distinct biological

relevance.

Low-frequency bands capture protein-scale features including overall charge distribution. Mid-frequency bands represent secondary structures and binding pockets. High-frequency bands isolate local features like catalytic atom-groups and coordinated metal centres. Computing separate Sinkhorn distances for each band produces a distance vector that preserves this scale separation.

This multi-scale approach revealed functional relationships invisible to single-metric comparisons. Charge changes associated with pH variation might present in lower bands while core functional elements such as catalysis can be expected to isolate in higher frequency bands. This theoretically allows isolation of different functional elements in different frequency bands. This capability exists only in this approach and potentially the spherical harmonic decomposition methods mentioned earlier in the chapter, but with the advantage of true three-dimensional comparison without the spatial dimensionality reduction of those methods.

Computational process

This section walks through the conceptual approach step-by-step as shown in Figure 5.13. The full technical details of the computational pipeline are found in Appendix A.

Point charge grid resolution and model uncertainty

The structural accuracy of AF2 models fundamentally limits the useful resolution for electrostatic field analysis. For the active site regions analysed here, which span approximately $18 \times 16 \times 20$ Å with a 10 Å padding buffer, the total volume reaches $38 \times 36 \times 40$ Å. The relationship between grid resolution and meaningful feature detection depends on the underlying model quality.

AF2 models with mean pLDDT scores around 90 represent very high confidence predictions, yet they still carry inherent positional uncertainties, particularly with amino acid side-chain positioning. The pLDDT metric predicts the local distance difference test (IDDT-C α), which measures how well local atomic distances are preserved rather than global RMSD. Validation against experimental structures indicates that regions with pLDDT scores above 90 typically achieve C α positional errors below 1 Å, while scores between 70 and 90 correspond to expected errors of 1 Å to 2 Å (Jumper et al., 2021; Mariani et al., 2013). Scores between 50 and 70 indicate low confidence with errors that may exceed 2 Å, and regions below 50 are often intrinsically disordered and should not be interpreted as reliable structure predictions.

The Nyquist-Shannon sampling theorem establishes that features smaller

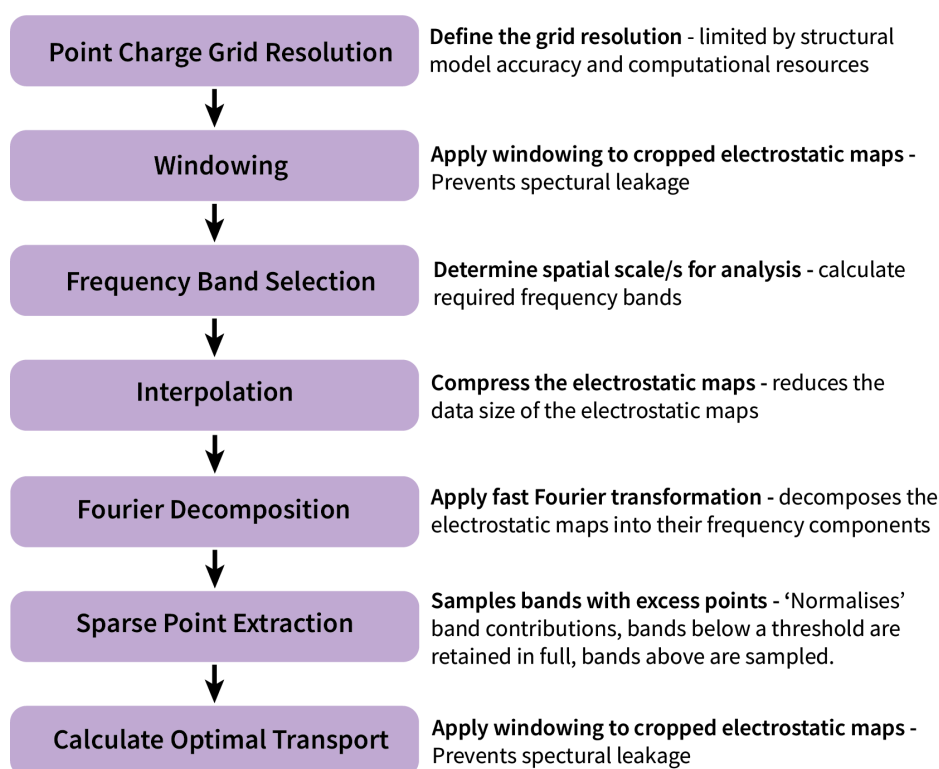


Figure 5.13: Flow chart of the core computational processes for Approach 3, additional to the structural modelling and active site cropping from Approaches 1 and 2. Each major step (purple box) is detailed in the text.

than approximately twice the grid spacing cannot be reliably distinguished. For the chosen 64^3 grid over the active site volume, the spacing averages approximately $0.59 \text{ \AA}$ (specifically $0.59 \text{ \AA}$, $0.56 \text{ \AA}$, and $0.63 \text{ \AA}$ in the three directions). This resolution captures atomic-level electrostatic features including individual atom positions (van der Waals radii typically range from $1.5 \text{ \AA}$ to $2.0 \text{ \AA}$), hydrogen bonding geometries (typically $1.8 \text{ \AA}$ to $2.2 \text{ \AA}$), and close atomic contacts between residues. While this resolution falls within the accuracy of the AF2 models at a glance, the issue is side chain positioning. The AF2 accuracy is for backbone elements only, and not those of the side chains which can vary in length from $0 \text{ \AA}$ for glycine to $7 \text{ \AA}$ for arginine. This error exceeds the resolution available for the 64^3 grid, although it is important to note that where side chain positioning has not been correctly predicted the actual error will be smaller than the length of the side chain, as it will still be sterically constrained, particularly given the close confines of the active site.

Increasing the grid resolution to 96^3 yields an average spacing of $0.40 \text{ \AA}$, while a 128^3 grid would provide approximately $0.30 \text{ \AA}$ spacing. These finer resolutions exceed the positional accuracy of the underlying AF2 models. When the grid spacing becomes substantially smaller than the model's backbone positional uncertainty of $1 \text{ \AA}$ to $2 \text{ \AA}$, the additional resolution captures variations

in the model that do not reflect genuine structural differences but rather modeling uncertainty and noise. This false precision not only wastes computational resources but may also introduce spurious differences between proteins that arise from model uncertainty rather than true electrostatic variations.

The pLDDT metric itself predicts local distance accuracy rather than global positioning. High pLDDT regions can still be positioned incorrectly relative to each other, and the metric tends toward overconfidence in some cases. Side chain positions carry more uncertainty than backbone atoms even in high-confidence regions. These limitations reinforce that grid resolutions finer than the model's inherent accuracy provide no meaningful additional information for electrostatic field comparison.

For the highest frequency band analysed here, ranging from $0.35 \text{ \AA}^{-1}$ to $0.45 \text{ \AA}^{-1}$ (corresponding to wavelengths of $2.2 \text{ \AA}$ to $2.9 \text{ \AA}$), the 64^3 grid with $0.59 \text{ \AA}$ spacing provides a Nyquist frequency of $1/(2 \times 0.59) = 0.85 \text{ \AA}^{-1}$, safely above the highest frequency of interest. This demonstrates that the 64^3 grid provides sufficient resolution to capture all biologically relevant electrostatic features but operates near the meaningful precision boundaries set by the AF2 model accuracy with regard to side chains. Accordingly, we would expect to see more noise in the higher frequency bands.

Window function application

The artificial boundaries created by cropping electrostatic maps to the active site region introduce a significant challenge for Fourier analysis. These create sharp transitions from the electrostatic field values to zero at the map edges. This results in high-frequency artefacts in the Fourier transform, a phenomenon known as spectral leakage. These artefacts can obscure genuine frequency components and limit comparison between proteins.

To mitigate these edge effects, the implementation applies a three-dimensional Hann window (Harris, 1978) to the interpolated data before Fourier transformation. The Hann window was selected as it provides a good balance between frequency resolution and spectral leakage suppression. Alternative windows such as Blackman (Harris, 1978) offer better leakage suppression but at the cost of a wider main lobe and reduced frequency resolution, while rectangular windows cause severe spectral leakage. The Hann window requires no parameter tuning, unlike Kaiser windows (Harris, 1978), making it a robust default choice for this application.

The Hann window smoothly tapers the electrostatic field from its full value in the centre of the grid to zero at all boundaries (Appendix A). This gradual transition eliminates the sharp discontinuities that cause spectral leakage, though

at the cost of some information loss near the map boundaries.

This windowing strategy requires careful consideration of the interaction between padding distance and window size. The implementation adds padding around each active site to accommodate the full extent of electrostatic features, but then applies the Hann window across the entire padded grid. The window function weights drop to 0.5 at approximately one-quarter of the distance from the edge and approach zero at the boundaries. This ensures smooth transitions that prevent spectral leakage but may inadvertently attenuate important electrostatic features located away from the grid centre.

The impact of this windowing approach can be assessed by visualising the frequency filtered fields using MIP. This technique reduces the three-dimensional electrostatic field to a two-dimensional representation by selecting the maximum potential magnitude at each lateral position across all depths. The resulting visualisation reveals the spatial extent and distribution of electrostatic features in each frequency band.

Figure 5.12C shows the frequency decomposition of AtAHL at pH 8.2 into three bands applying MIP. The extensive regions surrounding the active site with near-zero potential indicate that the Hann window may be excessive and attenuates any electrostatic features that extend beyond the central region. This effect disproportionately impacts the lowest frequency band (>10 Å), as these long-wavelength features require larger spatial extents to be properly characterised. A single oscillation of a 10–20 Å wavelength feature requires most of the grid space, yet the window function suppresses these peripheral regions where such features would be found. Consequently, long-range electrostatic steering fields may be severely underestimated.

Frequency domain analysis The Fourier transformation converts the spatial electrostatic field into a frequency representation that reveals features at different length scales and the concept has been described above. The technical implementation is described in Appendix A.

Frequency band selection The periodicity of the patterns in the electrostatic maps are represented by a continuous frequency spectrum. Initially the spectrum was broken up into six discrete bands, each corresponding to distinct structural features observed in protein structures. This provided a logical basis for the band cutoffs; although since Fourier transformations haven't been applied in this way before this was speculative. This division transforms the complex three-dimensional frequency data into manageable components that can be individually analysed and compared. These band boundaries were re-

fined through testing with 15 bands, which were then consolidated back into 6 non-uniform bands, distinct from the initial test, that balance information content with computational efficiency. This consolidation reduced computational overhead by a factor of 2.5 while preserving discriminatory power by grouping highly correlated frequency ranges and isolating non-correlated bands. Of note, later I was able to improve the code to calculate all bands simultaneously (see the computational pipeline section). However, for this proof-of-concept study I chose to still limit the number of bands tested to six because interpretation and downstream analysis becomes unwieldy as the number of bands increases. These bands are described below.

Band 1 captures the lowest frequencies, $f \in [0.05, 0.10) \text{ \AA}^{-1}$, corresponding to wavelengths $\lambda = 1/f \in (10.0, 20.0] \text{ \AA}$. These long wavelengths capture global fold organisation and domain-level electrostatic features. Changes in this band reflect large-scale differences in protein architecture or the overall distribution of charge across the protein.

Band 2 covers intermediate frequencies, $f \in [0.10, 0.15) \text{ \AA}^{-1}$, with wavelengths $\lambda \in (6.7, 10.0] \text{ \AA}$. This band covers features including spacing of secondary structure and typical binding pocket dimensions. Distances between maps at this scale suggest differences in alpha helices and beta sheets or in the size and shape of active site cavities.

Band 3 focuses on frequencies $f \in [0.15, 0.20) \text{ \AA}^{-1}$, where wavelengths are $\lambda \in (5.0, 6.7] \text{ \AA}$. This range is at a scale that is known to be the characteristic pitch of alpha helices at approximately 5.4 $\text{\AA}$.

Band 4 targets $f \in [0.20, 0.27) \text{ \AA}^{-1}$, corresponding to wavelengths $\lambda \in (3.7, 5.0] \text{ \AA}$. These wavelengths match the typical spacing between beta strands at approximately 4.7 $\text{\AA}$. This band is likely particularly sensitive to changes in beta sheet organisation and the presence of extended structures.

Band 5 examines higher frequencies, $f \in [0.27, 0.35) \text{ \AA}^{-1}$, with wavelengths $\lambda \in (2.9, 3.7] \text{ \AA}$. This range includes hydrogen bonding distances. Differences in this band often reflect changes in the detailed interaction networks within the active site.

Band 6 represents the finest scale features, $f \in [0.35, 0.45) \text{ \AA}^{-1}$, corresponding to wavelengths $\lambda \in (2.2, 2.9] \text{ \AA}$. This band captures close contacts at approximately 2.3 $\text{\AA}$. While likely at the resolution limit of the method given system noise, this band can distinguish proteins with different patterns of tight atomic packing or covalent cross-links.

Interpolation

The analysis takes as input the preprocessed electrostatic maps generated by APBS in OpenDX format. These files contain electrostatic potential values in units of kT/e on regular three-dimensional grids that often exceed 800,000 voxels. Processing such large datasets directly would be computationally prohibitive and unnecessary, as most voxels contain negligible electrostatic potential values.

The implementation extracts only electrostatically significant points where the absolute potential exceeds 10^{-10} kT/e, typically yielding between 8,000 and 12,000 significant points per protein structure and achieving approximately 98 % compression of the original data. When extraction yields more than 12,000 significant points, the method employs probabilistic importance sampling weighted by absolute potential magnitude, ensuring that regions with stronger electrostatic fields are preferentially retained in the sparse representation (Appendix A). This is analogous to compressing a photo on a computer so it takes up less memory. Points with weaker potentials are more likely to be excluded when downsampling is necessary; this limitation is addressed in the Discussion (5.4.6).

Fourier transformation The Fast Fourier Transformation algorithm is applied to each electrostatic map. For our active site region measuring $18 \times 16 \times 20$ Å with a 10 Å padding buffer, the total volume reaches $38 \times 36 \times 40$ Å. With $n = 64$, this yields frequency resolutions of:

- $\Delta f_x \approx 0.026 \text{ Å}^{-1}$,
- $\Delta f_y \approx 0.028 \text{ Å}^{-1}$, and
- $\Delta f_z = 0.025 \text{ Å}^{-1}$.

The maximum frequencies are:

- $f_{\max}^x \approx 0.84 \text{ Å}^{-1}$,
- $f_{\max}^y \approx 0.89 \text{ Å}^{-1}$, and
- $f_{\max}^z = 0.80 \text{ Å}^{-1}$.

These parameters provide sufficient resolution to distinguish features ranging from the overall active site architecture down to individual atomic contacts³.

Sparse field extraction After interpolation to a regular grid and Fourier transformation, a second stage of sparse extraction occurs in frequency space. Points within each frequency band are ranked by power magnitude, and when

³Note that while these parameters can resolve down to atomic level contacts, this is false precision exceeding the error of the underlying structural models.

a band exceeds 1,000 points above the noise threshold, the top 1,000 by power are selected deterministically; bands with fewer points retain all available data. The resulting point cloud (X_b, w_b) uses normalised power as weights:

$$w_b^{(i)} = \frac{P_{\text{masked}}^{(b)}(\mathbf{k}_i)}{\sum_{j=1}^N P_{\text{masked}}^{(b)}(\mathbf{k}_j)} \quad (5.16)$$

This deterministic top- k selection in frequency space ensures reproducibility and captures the dominant spectral features within each band.

Optimal transport The pilot test (approach 1) used [EMD](#) to calculate distances between the electrostatic maps, however [EMD](#) is computationally expensive and does not scale well, as it requires solving a linear programming problem. The computational cost becomes prohibitive for datasets with thousands of points such as in our electrostatic maps.

The Sinkhorn algorithm addresses this by reformulating the optimal transport problem. Regularisation is a technique that modifies an optimisation problem to make it easier to solve, typically by adding a penalty term that constrains the solution space. In the Sinkhorn algorithm, an entropy regularisation term is added to the optimal transport function. Entropy, in this context, measures how “spread out” the solution space is. A plan with high entropy distributes mass across a large solution space, while a plan with low entropy concentrates mass on a smaller space.

The regularisation term penalises transport plans with low entropy, whether those that are too sparse or concentrated. This has the effect of the algorithm finding a “smoothed”, or approximated solution.

This reformulation transforms the problem from one requiring linear programming into one solvable through iterative matrix operations which is much more efficient and highly parallelisable. The algorithm iterates until the changes between iterations fall below a convergence threshold.

The trade-off is that entropy regularisation introduces a bias where the solution is no longer the exact minimum cost but an approximation. The regularisation parameter ϵ controls this trade-off with smaller values producing solutions closer to the true [EMD](#) but requiring more iterations to converge, while larger values converge faster but with looser approximations. For distribution comparison such as between electrostatic maps this slight bias is typically acceptable given the computational gains and can even be advantageous. Small perturbations in the input distributions can cause unexpected “jumps” in value, essentially adding more noise to the noise. The entropy regularisation in Sinkhorn prevents these discontinuous jumps, smoothing out the noise. For pair-

wise comparisons of enzyme electrostatic maps, this smoothness means the distance metric responds predictably to small variations in the input, making it more robust for similarity assessment between closely related enzyme structures, mitigating some of the accuracy limitations of the AF2 models. Further detail can be found in Appendix A.

Parameter optimisation

While computational efficiency had to be optimised to ensure practical run times, the biological analysis required optimisation of the method parameters.

With approximately 10 tunable parameters – many were confounded – in this approach, each with at least a minimum average of four different values that could be set, the potential optimisation space exceeds $4^{10} = 1,048,576$ different combinations. As such, I chose to optimise only key parameters and do so individually.

SINKHORN CONVERGENCE PARAMETERS

The Sinkhorn algorithm's convergence behaviour depends critically on the regularisation parameter ϵ and the convergence threshold τ . Testing explored epsilon multipliers from 0.01 to 10.0.

The epsilon multiplier determines how the regularisation scales with the characteristic distance between point clouds. Pilot tests revealed that $\alpha = 1.00$ to 0.01 provides robust convergence across all protein pairs. Lower values such as $\alpha = 0.001$ risked convergence failures, particularly for proteins with very different charge distributions. This means the maximum number of iterations has to be increased proportionally which significantly increases run time. Higher values like $\alpha = 10.0$ remained stable but can produce over-smoothed transport plans that reduced discriminative power or may essentially reduced the optimal transport to a simple weighted average, losing all spatial information. The value α that was used experimentally was determined in conjunction with the grid resolution.

GRID RESOLUTION

The interpolation grid resolution fundamentally limits the spatial frequencies that can be resolved. Pilot tests across resolutions from 48^3 to 128^3 revealed clear trade-offs between computational cost and classification accuracy.

Initial testing showed that the 48^3 resolution had reduced ability to discriminate distance. Higher resolutions of 96^3 and 128^3 increased computation time by factors of 1.2 and 2 respectively, with minimal improvement in classifica-

tion accuracy. The 128^3 resolution also encountered memory limitations on available hardware, requiring additional batch processing that further slowed computation.

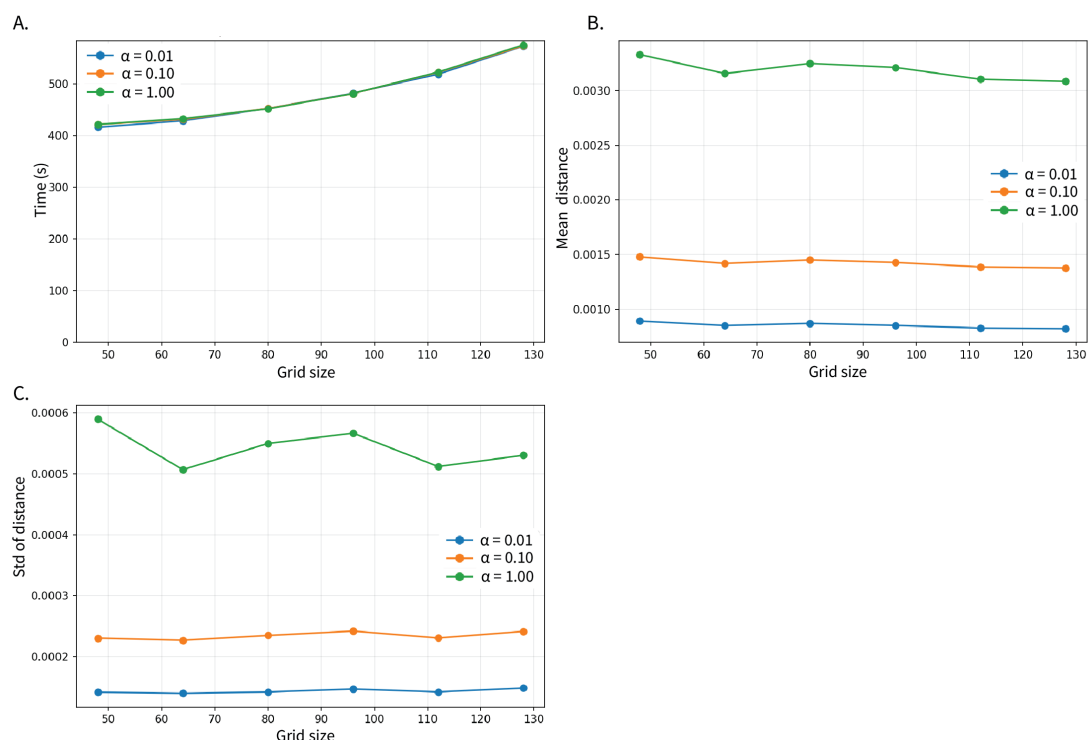


Figure 5.14: Electrostatic map grid size parameter sweep against the epsilon (regularisation term) multiplier α for 100 randomised electrostatic active site maps. α values tested of 0.01 (blue), 0.10 (orange), 1.00 (green) (A) Computational time by grid size tested against the α values. (B) Mean Sinkhorn distance. (C) Standard deviation of Sinkhorn distance.

With these broad bounds defined, I ran a parameter sweep using a subset of 100 randomly selected electrostatic maps. I tested three values of the ϵ multiplier α (0.01, 0.10, 1.00) against six grid sizes (48^3 , 64^3 , 80^3 , 96^3 , 112^3 , and 128^3), comparing run time, mean Sinkhorn distance, and standard deviation of Sinkhorn distance (Figure 5.14). As expected, run time increased with grid size (Figure 5.14A). The time increase appears modest in this test because all 100 electrostatic maps could be loaded onto the GPU in a single batch. The full dataset requires 595 batches at a grid size of 64^3 , and batch operations become increasingly costly at larger grid sizes. This creates a multiplicative effect on computation time that is not fully captured in Figure 5.14A. Notably, the three α values did not meaningfully affect run time.

Figure 5.14B shows that higher α values produced larger mean distances between electrostatic maps. This reflects the mathematical behaviour of the Sinkhorn algorithm as larger ϵ values introduce more entropic regularisation, which blurs the optimal transport plan and inflates distance estimates. Lower epsilon values preserve finer distinctions between distributions. Consistent

with this, separate testing on paralogue clustering confirmed that $\alpha = 0.01$ provided better discrimination between paralogues than $\alpha = 1.0$ via higher R^2 variance, despite producing smaller absolute distances (Table 5.2).

Table 5.2: PERMANOVA R^2 Band 4 and pH condition for paralogue classification at different ϵ multiplier values.

ϵ multiplier	pH7.0	pH7.3	pH7.6	pH7.9	pH8.2
0.01	0.326***	0.327***	0.333***	0.334***	0.335***
1.00	0.259***	0.258***	0.261***	0.263***	0.258***

Significance levels: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Adj-Sig indicates adjusted significance after false discovery rate correction.

Figure 5.14C shows that standard deviation of distance remained relatively stable across grid sizes for all α values. Distance metrics showed little variation between grid sizes of 64^3 and 128^3 , suggesting diminishing returns at higher resolutions. The slight increase in both mean distance and standard deviation between 64^3 and 80^3 is consistent with approaching the noise floor, where additional grid resolution captures noise rather than signal, as discussed earlier. However, these differences are subtle and the sample size of 100 maps limits the strength of this conclusion.

Given the stability of distance metrics above 64^3 and the substantial computational cost of higher resolutions, I selected 64^3 as the working resolution. This gives a frequency resolution of approximately $0.0156 \text{ \AA}^{-1}$ and a grid spacing of roughly $1.5 \text{ \AA}$, which is similar to the typical resolution of AF2 structures. According to the Nyquist criterion, this grid spacing can reliably capture electrostatic features with spatial wavelengths down to approximately $3 \text{ \AA}$. Features smaller than this may be under-sampled. However, frequency band testing (see Frequency band optimisation below) showed that wavelengths below $2.2 \text{ \AA}$ captured predominantly noise rather than biologically meaningful signal. The 64^3 resolution therefore adequately samples the frequency range containing useful information.

POINTS PER BAND OPTIMISATION

The number of frequency points extracted from each band affects both computational cost and the quality of the optimal transport solution. Testing with 500, 1,000, and 2,000 points per band revealed that 1,000 points provides the optimal balance.

With only 500 points, the optimal transport algorithm produced unstable solutions for bands with broadly distributed frequency content. The sparse sampling inadequately represented the continuous frequency distribution, leading to transport plans that varied significantly with small perturbations in the input data.

Extracting 1,000 points per band captured the majority of the spectral energy while ensuring stable convergence of the Sinkhorn algorithm. This density provides sufficient resolution to distinguish subtle differences in frequency distributions without overwhelming computational resources.

Increasing to 2,000 points doubled computation time with negligible improvement in classification accuracy in qualitative pilot testing. The additional points primarily captured low-power frequency components that contributed minimally to the distance calculations. These lower frequency components equate to band 1 or lower frequencies, therefore measuring at scales of $>10\text{--}20\text{ \AA}$ which is already at the same scale as the active site itself ($18 \times 16 \times 20$). Increasing this will introduce artefacts through insufficient sampling at this scale.

FREQUENCY BAND OPTIMISATION

Initial analysis employed 15 frequency bands spanning spatial scales from $20.0\text{ \AA}$ to $1.5\text{ \AA}$ (Figure 5.15). These bands were defined with increasingly narrow spatial ranges at higher frequencies. This non-uniform binning reflects the inverse relationship between frequency and spatial scale in Fourier analysis where equal intervals in frequency space correspond to progressively narrower intervals in spatial (wavelength) space as scale decreases. Additionally, finer spatial scales contain more structural detail, requiring higher resolution binning to adequately capture variation in electrostatic features at these scales.

15 Band spatial scales:

- **Band 1:** $11.1\text{--}20.0\text{ \AA}$
- **Band 2:** $7.7\text{--}11.1\text{ \AA}$
- **Band 3:** $5.9\text{--}7.7\text{ \AA}$
- **Band 4:** $4.8\text{--}5.9\text{ \AA}$
- **Band 5:** $4.0\text{--}4.8\text{ \AA}$
- **Band 6:** $3.4\text{--}4.0\text{ \AA}$

- **Band 7:** 3.0–3.4 Å
- **Band 8:** 2.7–3.0 Å
- **Band 9:** 2.4–2.7 Å
- **Band 10:** 2.2–2.4 Å
- **Band 11:** 2.0–2.2 Å
- **Band 12:** 1.9–2.0 Å
- **Band 13:** 1.8–1.9 Å
- **Band 14:** 1.6–1.8 Å
- **Band 15:** 1.5–1.6 Å

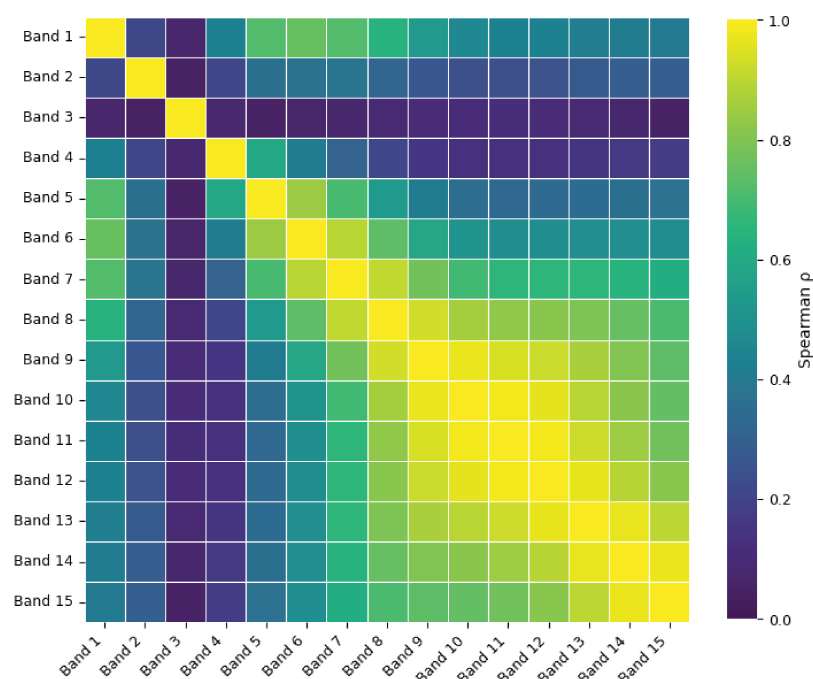


Figure 5.15: Spearman correlation matrix between frequency band electrostatic similarities and total Fourier distance. Bands 1–15 represent spatial frequency components from coarse (Band 1: 11.1–20.0 Å) to fine (Band 15: 1.5–1.6 Å) scales. Fourier distance represents the aggregate similarity metric across all frequency bands. High correlations between adjacent fine-scale bands (Bands 9–15) indicate redundant structural information, while coarse-scale bands show more independent variation.

To reduce the 15 bands to a more analytically tractable number while retaining maximal information, I examined inter-band correlations to identify which frequency ranges contributed independent information to the analysis. The correlation matrix (Figure 5.15) revealed distinct patterns across the frequency spectrum. Fine-scale bands (Bands 9–15, corresponding to spatial scales below 2.7 Å) showed high inter-correlation, indicating substantial redundancy in structural information at these scales. In contrast, Bands 3 and 4 (5.9–7.7 Å and 4.8–5.9 Å respectively) showed the lowest correlations with other bands, suggesting that the mid-frequency range spanning approximately 4.8–7.7 Å captures electrostatic features that are most distinct from both coarser and finer

spatial scales.

Based on this analysis, I consolidated the 15 bands into 6 bands optimised to maximise independent information content. Bands corresponding to spatial scales below 2.2 Å (original Bands 11–15) were excluded entirely, as these frequencies were calculated to be disproportionately affected by structural and numerical noise; this is discussed in detail later in the chapter. The central range of the low-correlation region (the combined span of original Bands 3 and 4) was prioritised, with the 6-band configuration allocating disproportionately finer resolution to this mid-frequency region:

Partially optimised 6 Band spatial scales:

- **Band 1:** 10.0–20.0 Å
- **Band 2:** 6.7–10.0 Å
- **Band 3:** 5.0–6.7 Å
- **Band 4:** 3.7–5.0 Å
- **Band 5:** 2.9–3.7 Å
- **Band 6:** 2.2–2.9 Å

This allocation is evident in the band widths where Band 1 spans 10.0 Å, while the mid-frequency Bands 3 and 4 together span only 3.0 Å (5.0–6.7 Å and 3.7–5.0 Å), providing approximately three-fold higher resolution in the region identified as most informative. The remaining fine-scale bands (original Bands 7–10) were consolidated into Bands 5 and 6, reducing redundancy while retaining sensitivity to finer structural features above the 2.2 Å noise threshold. Further optimisation remains possible through systematic testing of the analysis pipeline at different bandwidths and boundary positions, particularly within the mid-frequency range where the correlation analysis suggests the greatest discriminatory power resides.

Biological assessment

This section details the biological results of Approach 3 in detail. These results are using the parameters that were approximately optimised. This optimisation process including the approach, challenges and limitations is discussed in detail in the preceding section 'Parameter optimisation'. In this section the Approach 3 results are broken down by feature/function. The first part of the section will provide the results of clustering by paralogue as a comparison between the other two approaches and the sequence-homology approach from chapter 4, this forms the basis of the approach's validation. The validation focused on whether the method could discriminate between paralogues as defined by the phylogeny in chapter 4 (Figure 4.3). Statistical comparison is not practical

between these different methods as the underlying data isn't equivalent including needing to use different distance metrics. Approach 1 and 2A used pairwise comparisons between all electrostatics maps and AtSAL1 at pH 7.9. Meanwhile sequence-homology, Approach 2B and Approach 3 all used all-versus-all comparisons. Sequence-homology is also unable to factor pH differences unlike Approaches 1-3. This challenge is most simply addressed through qualitative comparisons of similarity networks. While there are differences between the distance metrics and the data, we can still make visual comparisons of the clustering between these methods.

After having compared the different approaches, the results from Approach 3 will be explored in more detail to highlight opportunities and limitations of this method. It's ability to cluster based on taxa, pH and predicted localisation will be presented. Visual analysis will be presented and dendrogram heatmaps using euclidean distance and the average linkage method (UPGMA)⁴ as it is the standard method used in the python charting package used, and similarity networks which provide continuity from sequence-homology through to this approach. Statistical comparisons will be made using PERMANOVA. PERMANOVA was chosen as the primary statistical comparison method for Approach 3 as it allows multivariate analysis across multiple frequency bands, pH, and features (/functional groupings). All PERMANOVA analyses were run with number of permutations = 99,999 and minimum group size = 3.

VALIDATION

The Fourier and Sinkhorn parameters used in this validation were:

- Fourier grid resolution: 64^3 (262,144 voxels)
- Target points per band: 1,000
- Max points subsample: 12,000
- Sinkhorn threshold: 0.01
- ϵ multiplier: 0.01
- Max iterations: 50
- Min iterations: 5
- Sinkhorn LSE mode: True
- Frequency bands
 - Band 1: $(0.050, 0.100) \text{ \AA}^{-1} \mid \lambda = 10.0\text{--}20.0 \text{ \AA}$
 - Band 2: $(0.100, 0.150) \text{ \AA}^{-1} \mid \lambda = 6.7\text{--}10.0 \text{ \AA}$
 - Band 3: $(0.150, 0.200) \text{ \AA}^{-1} \mid \lambda = 5.0\text{--}6.7 \text{ \AA}$
 - Band 4: $(0.200, 0.270) \text{ \AA}^{-1} \mid \lambda = 3.7\text{--}5.0 \text{ \AA}$

⁴details can be found in the Methods chapter

- Band 5: $(0.270, 0.350) \text{ \AA}^{-1} \mid \lambda = 2.9\text{--}3.7 \text{ \AA}$
- Band 6: $(0.350, 0.450) \text{ \AA}^{-1} \mid \lambda = 2.2\text{--}2.9 \text{ \AA}$

PARALOGUE CLUSTERING

The quality of the results of this approach are immediately apparent. This dendrogram heatmap in Figure 5.16 shows that paralogue clustering is significantly improved over Approach 2 (Figure 5.9), with greater relative variation in distances. CNP is almost entirely contiguous, while AHL and SAL are largely contiguous with some mixing of the clusters. Fourier band four was chosen for this figure as the R^2 values for this band were the highest from PERMANOVA analysis presented in Table 5.3. At pH 8.2 $R^2 = 0.335$, where 33.5% of the variance in band 4 was explained by paralogue grouping with very high statistical significance (FDR corrected p-value <0.001).

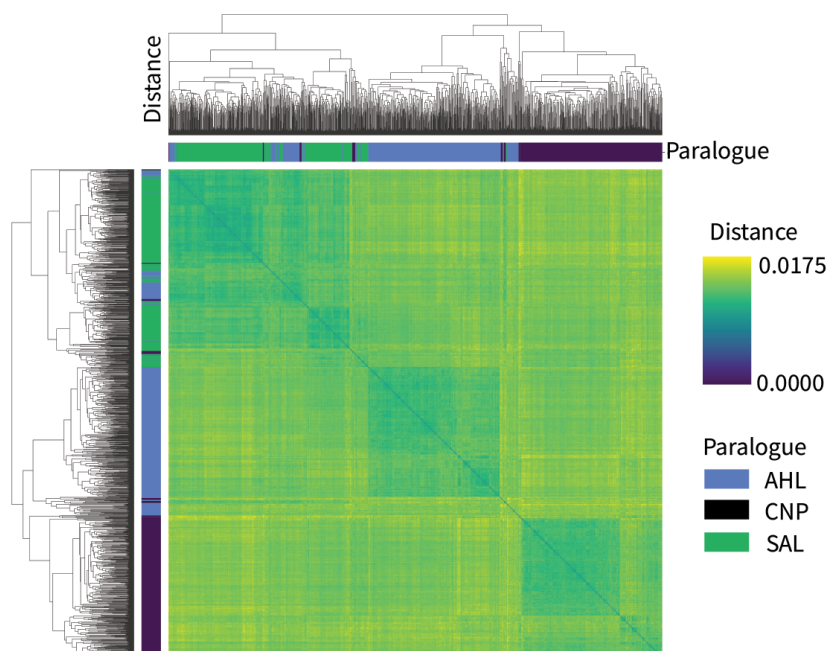


Figure 5.16: Fourier decomposition dendrogram heatmap of frequency band 4 (scale 3.7–5.0 $\text{\AA}$) with annotation by paralogue. The 2,530 electrostatic maps of all 506 enzymes at pH 7.0, 7.3, 7.6, 7.9, and 8.2 are represented. The heatmap is coloured by Sinkhorn distance. Dendrogram branch lengths are calculated from the Sinkhorn distance. Dendrogram is clustered using hierarchical clustering, see the Methods section.

When visualised as a similarity network in Figure 5.17 we can see the two-dimensional relationship between the clusters. In Figure 5.17 A. we can see CNP is in a distinct cluster with minimal mixing. SAL largely remains as a distinct cluster although some AHL nodes do cluster with it. AHL is fragmented into a main cluster with a sub-cluster localised with SAL. When we colour AHL by its own paralogues we see that AHL- α and AHL- β are mixed together, Basal AHL is partially mixed in with AHL- α and AHL- β , while Algal AHL forms the

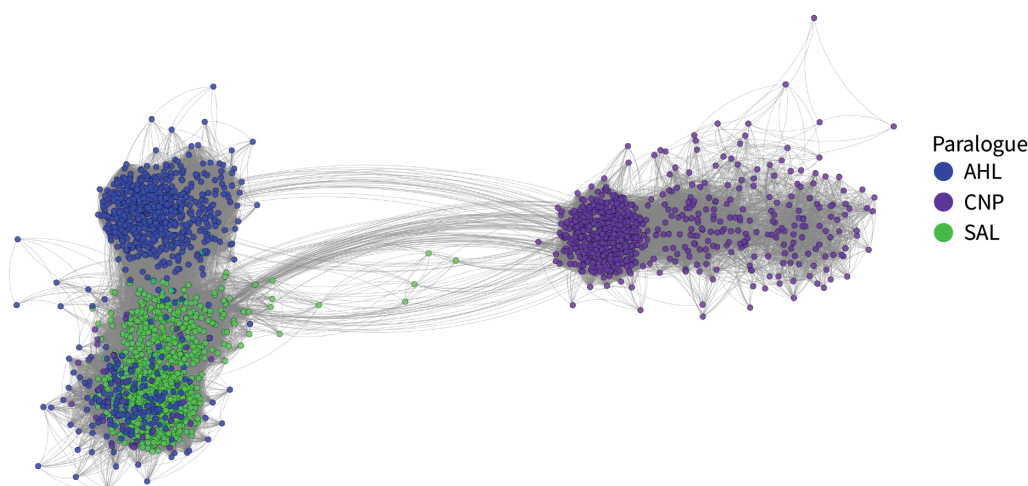
Table 5.3: PERMANOVA R^2 values by band and pH condition for paralogue classification.

Frequency band	pH7.0	pH7.3	pH7.6	pH7.9	pH8.2
Band 1	0.070***	0.086***	0.103***	0.118***	0.098***
Band 2	0.083***	0.082***	0.080***	0.085***	0.073***
Band 3	0.326***	0.327***	0.330***	0.326***	0.328***
Band 4	0.326***	0.327***	0.333***	0.334***	0.335***
Band 5	0.276***	0.275***	0.276***	0.274***	0.275***
Band 6	0.276***	0.279***	0.285***	0.284***	0.284***

Significance levels: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Adj-Sig indicates adjusted significance after false discovery rate correction.

cluster that is adjacent to SAL.

A.



B.

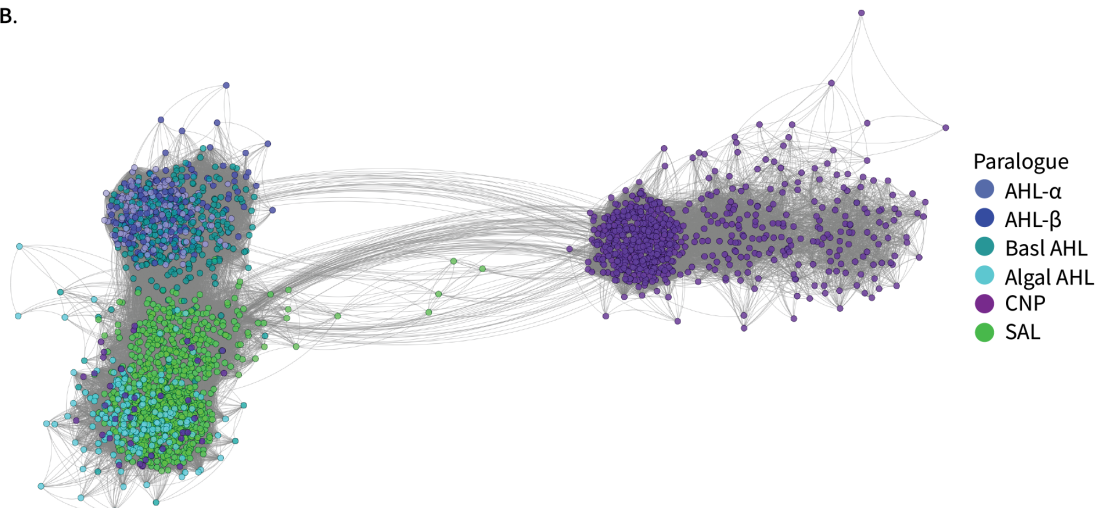


Figure 5.17: Fourier decomposed, frequency band 4 (scale 3.7–5.0 Å) active site electrostatic similarity network with Sinkhorn distances. Nodes represent frequency band 4 of decomposed active site electrostatic maps, with 2,530 node total of all 506 enzymes at pH 7.0, 7.3, 7.6, 7.9, and 8.2. Visualised in Cytoscape using Prefuse Force Directed Layout with edges filtered by the 0.08 quantile, coloured by paralogue as determined by phylogenetic analysis (see Figure 4.3). **A.** Paralogues shown are AHL: blue, CNP: purple, SAL: green. **B.** Paralogues shown are AHL- α : sky blue, AHL- β : blue, Basal AHL: sea green, Algal AHL: light blue, CNP: purple, SAL: green.

CONTROLS

Three core controls were used to ensure the data was not random or caused by artefacts. The first approach was a qualitative comparison between similarity networks generated by sequence and structural homology, and approach 3. The second control method was progressively adding random noise to the data and observing the degradation in ability to cluster by paralogue. The final control was paralogue distance comparison where labels were scrambled to identify whether or not clustering could be explained by randomness or artefacts.

To validate that the Fourier decomposition method provides information beyond sequence or structural similarity alone, comparison networks were constructed using alternative methods. FoldSeek ([van Kempen et al., 2024](#)) was used to generate a structural similarity network based on three-dimensional structure alignment scores. Foldseek represents each amino acid by both its identity and its local 3D interaction pattern with nearby residues. This approach creates a structural alphabet called 3Di that captures backbone geometry and residue-residue interactions. The method encodes protein structures as 3Di sequences, which enables Foldseek to apply fast sequence alignment algorithms similar to BLAST. This avoids the computational expense of traditional 3D structural superposition methods. Foldseek combines the amino acid sequence with the 3Di structural sequence to create a dual representation. This allows searches to identify proteins based on sequence similarity, structural similarity, or both characteristics simultaneously. It was chosen due to its speed and its ability to work with structural models while providing a simple similarity metric analogous to a distance score. This allowed me to create a structural similarity network. Figure 5.18 shows similarity networks based off A. sequence-homology, B. structural alignment, and C. Fourier transformed active sites.

Qualitative comparison of the similarity networks reveals that the Fourier decomposition method produces distinct clustering patterns compared to structural and sequence similarity approaches. While it is not possible to make direct statistical comparisons between the methods due to their fundamentally different distance metrics and that the sequence and structural networks do not include multiple comparisons based on pH, the network visualisations demonstrate that electrostatic similarity can distinguish between structurally similar proteins that have diverged in their electrostatic properties.

Of note, the distance between paralogues (AHL, CNP, SAL) is smallest in sequence homology yet it has greater separation with AHL's paralogues (Figure 5.18). Figure 5.18C shows considerable overlap between SAL and Algal AHL, while CNP is very distant with much fewer edges linking it to the other

clusters indicating it is more dissimilar in active site electrostatic frequency space. The more mixed Basal AHL, AHL- α and AHL- β electrostatic maps suggest higher functional overlap than the sequence and structural homology suggests, as does the clustering of Algal-AHL almost entirely within SAL. All three subfigures in Figure 5.18 show CNP outliers that do not cluster with the rest of CNP, rather they cluster with or near the other homologues. Exploring this anomaly further, and colouring them consistently between the three subfigures (Figure 5.19) they are all the same proteins in each subfigure, and all come from Chromista, Rhodophyta, and Glaucophyta, the most ancient plants which formed the outgroup in the phylogeny (chapter 4, Figure 4.3).

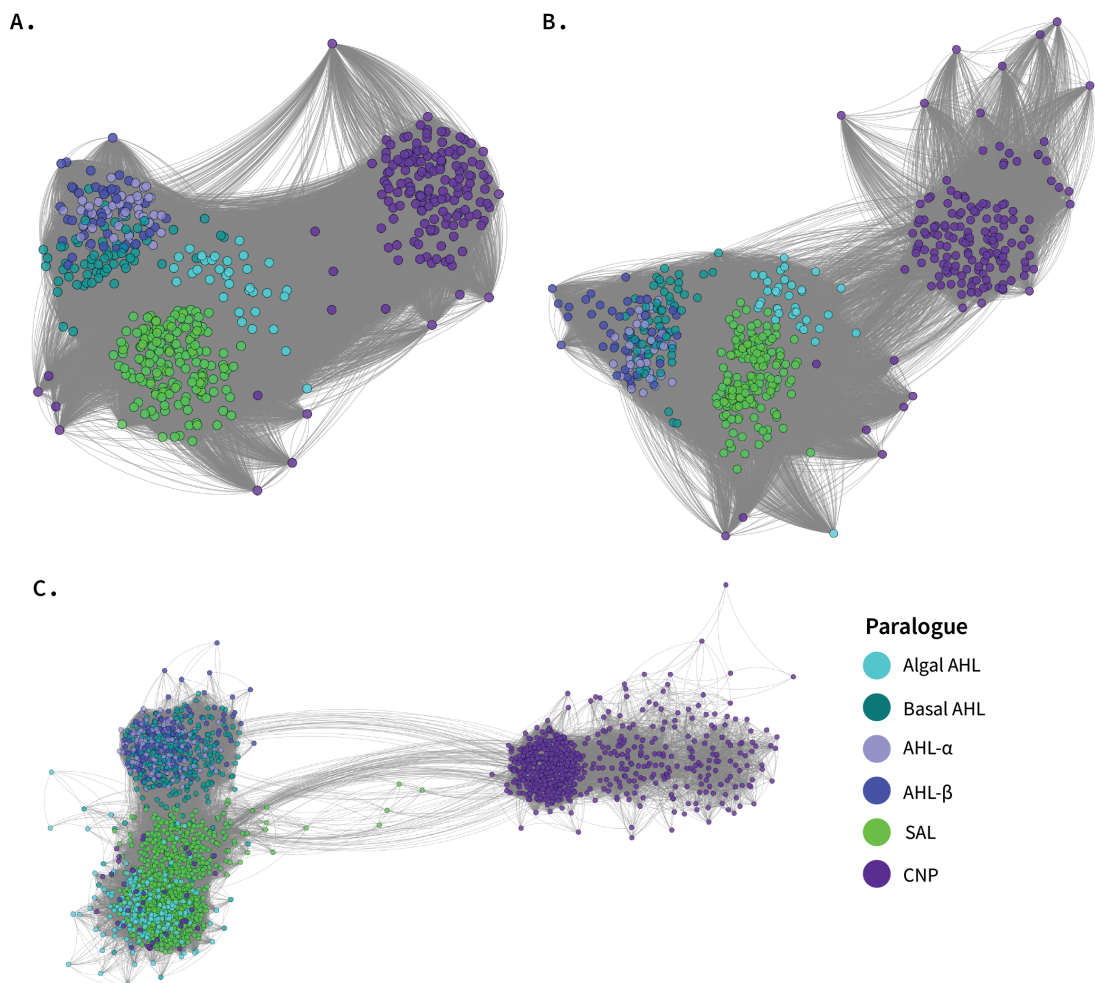


Figure 5.18: Similarity networks of the plant nucleotide phosphatases. A. Sequence similarity network where edges were filtered by Bit-score > 878. B. Structural similarity network where edges were filtered by Bit-score > 1,350. C. Fourier decomposition band 4 active site similarity network where edges were filtered to the 0.08 quantile. Paralogues shown are AHL- α : sky blue, AHL- β : blue, Basal AHL: sea green, Algal AHL: light blue, CNP: purple, SAL: green. Visualised in Cytoscape with Prefuse Force Directed Layout.

This suggests that the Fourier decomposition approach captures paralogue structure comparable to sequence and structural comparisons, though direct

quantitative comparison between methods was not possible due to different distance metrics. Figure 5.19 clearly demonstrates very similar data structure between the three comparison methods indicating that the data is not random, nor processing distinguishable artefacts. This is supported by testing using further controls; titratable insertion of noise into the data, and permutation analysis.

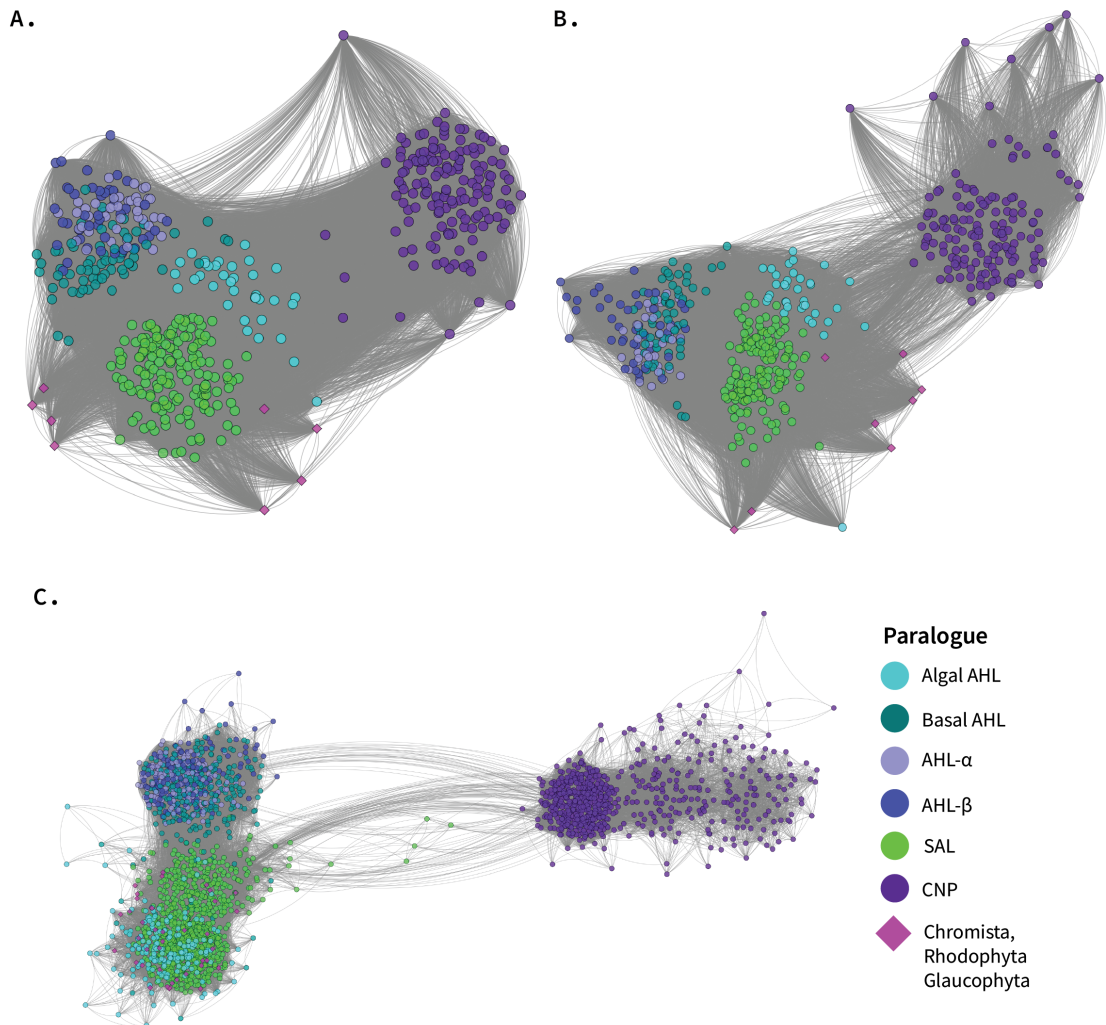


Figure 5.19: Similarity networks of the plant nucleotide phosphatases. A. Sequence similarity network where edges were filtered by Bit-score > 878. B. Structural similarity network where edges were filtered by Bit-score > 1,350. C. Fourier decomposition band 4 active site similarity network where edges were filtered to the 0.08 quantile. Paralogues shown are AHL- α : sky blue, AHL- β : blue, Basal AHL: sea green, Algal AHL: light blue, CNP: purple, SAL: green, Chromista, Rhodophyta, and Glaucophyta: pink diamonds. Visualised in Cytoscape with Prefuse Force Directed Layout.

Noise titration

Progressively adding noise to the data should degrade the clustering if the data is not random or an artefact. This can be statistically analysed.

Clustering robustness was evaluated through progressive addition of multiplicative noise to distance matrices. A log-normal noise model was employed

as it preserves relative distance relationships and accurately models proportional measurement errors.

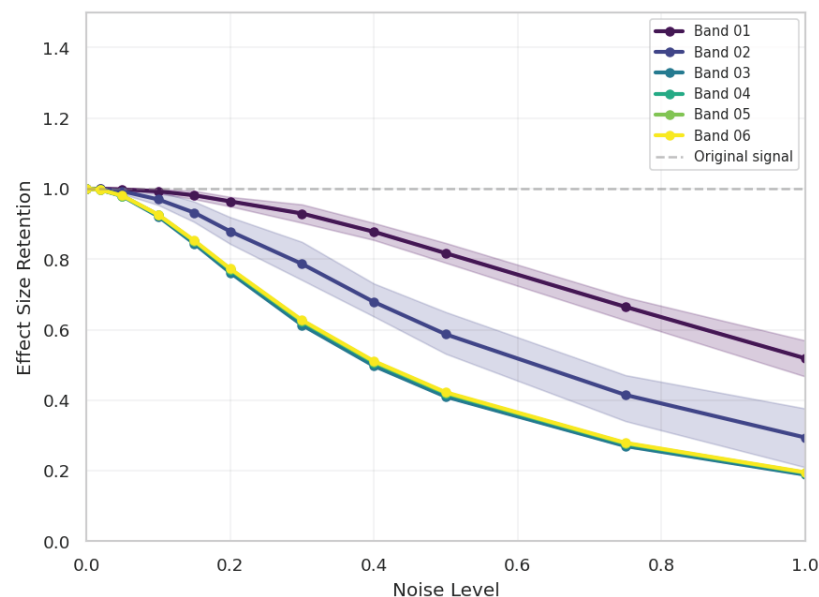


Figure 5.20: Progressive noise robustness analysis of hierarchical clustering across frequency bands. The **ARI**, a measure of clustering agreement between baseline and noise-perturbed clusters (ranging from -1 for complete disagreement to 1 for perfect agreement), was evaluated across increasing noise levels. Distance matrices were perturbed using multiplicative log-normal noise to simulate proportional measurement uncertainties while preserving relative distance relationships. For each band and noise level, clustering was performed with Ward's linkage hierarchical clustering over $n = 100$ iterations with different random seeds. Solid lines represent mean **Adjusted Rand Index (ARI)** values, while shaded regions indicate 95% confidence intervals derived from empirical percentiles.

Figure 5.20 shows a systematic decline in **ARI** (a measure of clustering agreement between baseline and noise-perturbed clustering) with increasing noise across all bands demonstrates that the observed clustering structure is not an artefact of random variation, with Band 1 showing least sensitivity to noise, followed by Band 2. Bands 3-6 have approximately the same decay patterns with very narrow 95% confidence intervals. At 100% noise Band 1 has an **ARI** of ~ 0.5 suggesting it contains $\sim 50\%$ noise. Continuing this logic Band 2 contains $\sim 30\%$ noise, while the other bands contain $\sim 20\%$ noise. This indicates that noise is most prominent in scales $>6.7 \text{ \AA}$.

Paralogue distance comparison (paralogue scrambling)

If the Fourier decomposition method captures biologically meaningful information about protein electrostatic fields, we would expect proteins that share evolutionary ancestry to have more similar electrostatic signatures than proteins from different lineages. This expectation arises because proteins within the same paralogue have inherited structural features and functional constraints

from a common ancestor, and these conserved characteristics should manifest in their electrostatic field patterns. To test this hypothesis, the average electrostatic distance between proteins from the same paralogue was compared against the average distance between proteins from different paralogues. If my method captures genuine evolutionary signal, within-paralogue distances should be systematically smaller than between-paralogue distances. To determine whether any observed difference exceeds what might occur by chance, a permutation approach was used where by randomly reassigning paralogue labels to proteins and recalculating the distance comparison, a null distribution was generated representing the expected difference when no true relationship exists between evolutionary history and electrostatic similarity. A significant result indicates that the frequency band captures evolutionarily conserved features of the electrostatic field at that spatial scale. Conversely, a non-significant result suggests the band may capture features that vary independently of evolutionary relationships, such as lineage-specific adaptations or methodological noise.

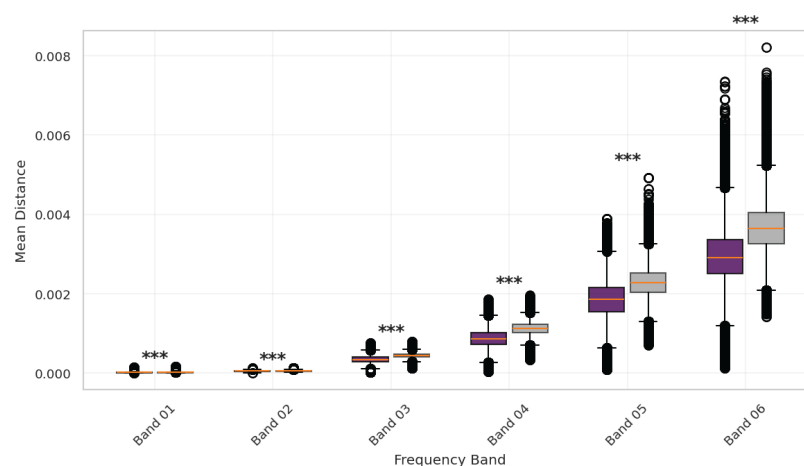


Figure 5.21: Parologue distance comparison control. Box plots show the distribution of pairwise electrostatic field distances classified as within-paralogue (purple) or between-paralogue (grey) for each of six frequency bands. Distances were calculated from 2,350 electrostatic maps (506 proteins $\times$ 5 pH conditions) pooled across three paralogues (AHL, CNP, SAL). Significance was assessed using a permutation test (1,000 iterations) comparing the observed difference between mean within- and between-paralogue distances against a null distribution generated by shuffling paralogue labels. *** $p < 0.001$. All frequency bands show significantly smaller within-paralogue distances, indicating that the Fourier decomposition captures evolutionarily conserved electrostatic field characteristics across all spatial scales examined.

Biological analysis beyond paralogues

TAXA

After demonstrating that the approach could successfully cluster paralogues, both statistically (Table 5.3) and qualitatively when compared to sequence and

structural homology methods (Figure 5.17), I tested whether it could capture other biologically meaningful patterns. Taxonomic grouping provided a logical next benchmark because the labels could be treated as ground truth. Since all sequences were derived from known species, taxonomic assignments could be assumed correct in the vast majority of cases, providing a reliable reference for assessing the approach's performance beyond paralogue relationships. Taxonomic order was used as the grouping variable for all analyses in this section.

Repeating the same PERMANOVA and similarity network analysis but using taxa as the grouping feature rather than paralogue revealed statistically significant clustering (Table 5.4, although the clustering was less cohesive compared to paralogue-based clustering (Table 5.3). All bands at all pH levels showed varying levels of statistical significance, with Band 1 having the lowest significance for variance explained by taxa. Band 3 had the highest R^2 values at $R^2 = 0.144 - 0.146$ between pH 7.6 and 8.2 with very high statistical significance (FDR corrected p-value < 0.001). This indicates that for the entire dataset, Band 3 was best able to cluster by taxa. Bands 1 and 2 performed poorly with regard to clustering power (R^2), a pattern that persisted across all subsequent analyses regardless of paralogue or pH condition.

Table 5.4: PERMANOVA R^2 values by band and pH condition for taxa classification.

Frequency band	pH7.0	pH7.3	pH7.6	pH7.9	pH8.2
Band 1	0.054*	0.058*	0.071**	0.091***	0.091***
Band 2	0.064***	0.061***	0.074***	0.069***	0.066***
Band 3	0.136***	0.137***	0.146***	0.146***	0.144***
Band 4	0.121***	0.121***	0.122***	0.123***	0.125***
Band 5	0.133***	0.133***	0.137***	0.136***	0.136***
Band 6	0.125***	0.123***	0.126***	0.123***	0.125***

Significance levels: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. FDR correction applied within each table independently.

Part of the reduced statistical support for clustering by taxa is that all three paralogues are found across Viridiplantae (see Chapter 4 Figure 4.3). This results in taxa classification being duplicated at least three times in most cases, unless a paralogue has been lost from a species. Paralogue and taxa features therefore compete for clustering power. Put another way, the data exhibit hierarchical structure with taxa nested within paralogue groupings, and both features contribute to the observed variance. A qualitative assessment aids in assessing clustering by taxa, which can be visualised in Figure 5.22.

The hierarchical structure can be partially disentangled by filtering by paralogue and then testing clustering by taxa. This approach only partially resolves the complication, as the electrostatic maps encode multiple biological features simultaneously. The preceding analysis established that both paralogue and

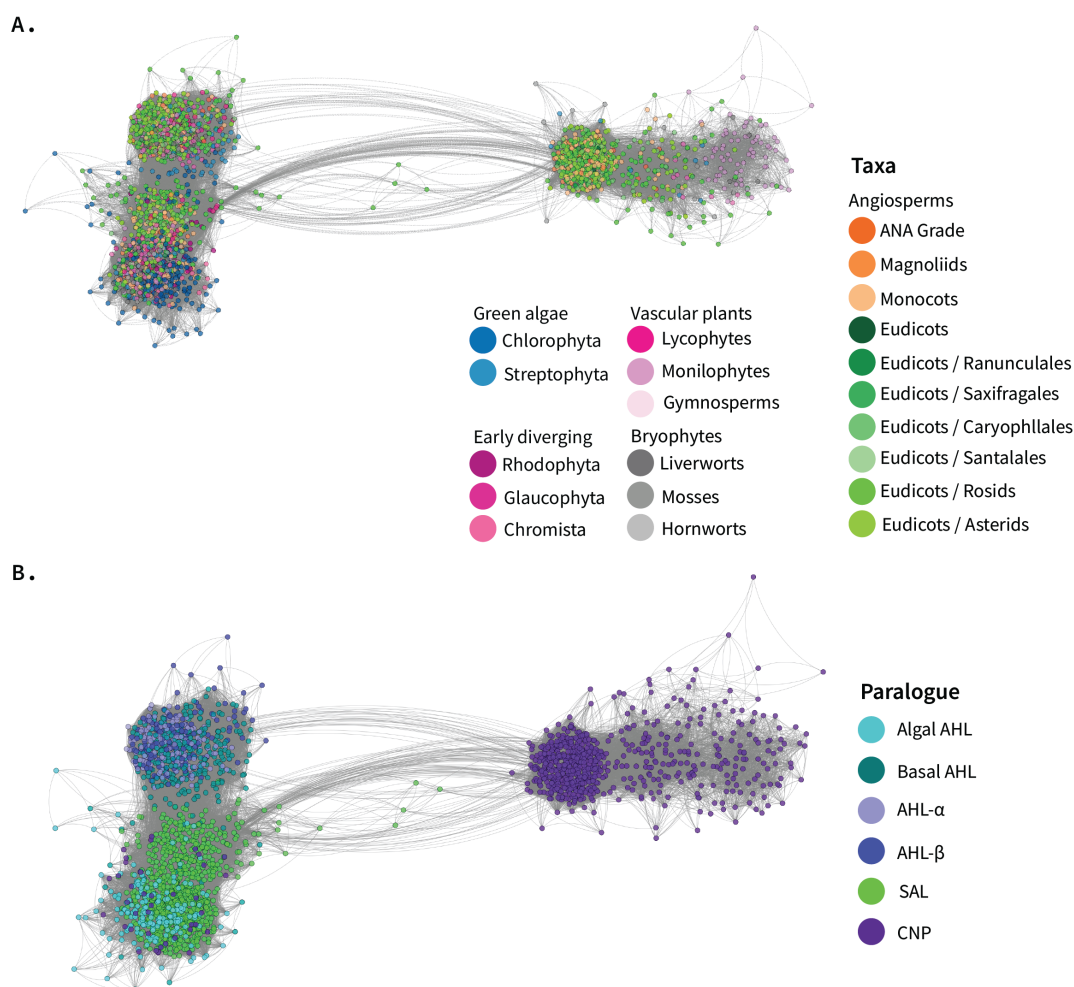


Figure 5.22: Fourier decomposed, frequency band 5 (scale 2.9–3.7 Å) active site electrostatic similarity network with Sinkhorn distances. Nodes represent frequency band 5 of decomposed active site electrostatic maps, with 2,530 nodes total of all 506 enzymes at pH 7.0, 7.3, 7.6, 7.9, and 8.2. Visualised in Cytoscape using Prefuse Force Directed Layout with edges filtered by the 0.08 quantile, coloured by taxa.

taxa information are encoded in the active site electrostatic maps. I tested this paralogue-filtered analysis with all three paralogues.

PERMANOVA analysis of AHL clustering by taxa ($n = 187$; Table 5.5) showed a substantial improvement in R^2 values compared to the full dataset. Band 3 showed the best ability to cluster by taxa, with R^2 values ranging from 0.304 to 0.312 across all pH conditions. The best R^2 of 0.312 at pH 7.6 represents more than a doubling of the clustering power compared to 0.146 for Band 3 in the full dataset (Table 5.4). Bands 1 and 2 continued to show weaker clustering, with Band 1 at pH 7.0 failing to reach statistical significance ($R^2 = 0.077$, ns).

PERMANOVA analysis of CNP clustering by taxa ($n = 159$; Table 5.6) also showed improved R^2 values compared to the full dataset. Bands 3, 4, and 5 performed similarly well, with R^2 values ranging from 0.253 to 0.260 at pH levels between 7.3 and 7.9. The best R^2 of 0.260 at pH 7.6 in Band 5 represents a

Table 5.5: AHL PERMANOVA R^2 values by band and pH condition for taxa classification (n = 187).

Frequency band	pH7.0	pH7.3	pH7.6	pH7.9	pH8.2
Band 1	0.077ns	0.104*	0.121**	0.166***	0.179***
Band 2	0.109***	0.116***	0.124***	0.122***	0.113***
Band 3	0.307***	0.307***	0.312***	0.305***	0.304***
Band 4	0.274***	0.272***	0.280***	0.277***	0.287***
Band 5	0.234***	0.231***	0.233***	0.234***	0.243***
Band 6	0.223***	0.224***	0.228***	0.231***	0.236***

Significance levels: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. FDR correction applied within each table independently.

substantial improvement over the full dataset maximum of 0.146. Bands 1 and 2 again showed weak clustering power, with Band 1 failing to achieve statistical significance at four of five pH conditions.

Table 5.6: CNP PERMANOVA R^2 values by band and pH condition for taxa classification (n = 159).

Frequency band	pH7.0	pH7.3	pH7.6	pH7.9	pH8.2
Band 1	0.090	0.085	0.080	0.097*	0.086
Band 2	0.091**	0.094**	0.091**	0.084*	0.075
Band 3	0.255***	0.259***	0.259***	0.254***	0.247***
Band 4	0.253***	0.257***	0.259***	0.257***	0.253***
Band 5	0.253***	0.257***	0.260***	0.256***	0.247***
Band 6	0.233***	0.230***	0.226***	0.223***	0.222***

Significance levels: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. FDR correction applied within each table independently.

PERMANOVA analysis of SAL clustering by taxa (n = 160; Table 5.7) similarly showed improved R^2 values compared to the full dataset. Band 3 at pH 8.2 showed the best ability to cluster by taxa, with the highest R^2 of 0.277. This represents a substantial improvement over the full dataset maximum of 0.146. Bands 4 and 5 also performed well, particularly at higher pH values.

Table 5.7: SAL PERMANOVA R^2 values by band and pH condition for taxa classification (n = 160).

Frequency band	pH7.0	pH7.3	pH7.6	pH7.9	pH8.2
Band 1	0.123*	0.112*	0.136*	0.156*	0.241***
Band 2	0.119**	0.111*	0.130***	0.102*	0.123***
Band 3	0.216***	0.216***	0.231***	0.240***	0.277***
Band 4	0.225***	0.227***	0.230***	0.233***	0.241***
Band 5	0.228***	0.231***	0.238***	0.242***	0.246***
Band 6	0.191***	0.189***	0.194***	0.202***	0.205***

Significance levels: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. FDR correction applied within each table independently.

Comparing across paralogues revealed several consistent patterns and notable differences. Filtering by paralogue increased the maximum R^2 values by

1.8 to 2.1 fold, from 0.146 in the full dataset to 0.260–0.312 in the paralogue-filtered analyses. This improvement confirms that the hierarchical structure was obscuring taxonomic signal in the combined analysis. Across all analyses, Bands 1 and 2 consistently underperformed regardless of paralogue or pH condition, and the mid-frequency bands (3–5) carried the strongest taxonomic signal.

Each paralogue exhibited a distinct optimal band profile for taxonomic clustering. AHL showed a clear dominance of Band 3, with R^2 values (0.304–0.312) distinctly separated from other bands. CNP showed a more distributed pattern, with Bands 3, 4, and 5 performing equivalently (R^2 approximately 0.253–0.260) and no single dominant band. SAL showed an intermediate pattern with clustering power distributed across Bands 3–5, and Band 3 performing best at pH 8.2.

The paralogues also differed markedly in their pH sensitivity. AHL Band 1 showed strong pH dependence, with R^2 more than doubling from 0.077 (not significant) at pH 7.0 to 0.179 ($p < 0.001$) at pH 8.2. SAL showed a similar pattern, with Band 1 R^2 increasing from 0.123 at pH 7.0 to 0.241 at pH 8.2, and Band 3 showing a notable increase at pH 8.2 (0.277) compared to pH 7.0 (0.216). In contrast, CNP showed remarkable stability across all pH conditions, with R^2 values varying minimally across the pH range tested.

PERMANOVA tests whether groups differ overall but does not identify which specific taxonomic orders cluster together or apart. Pairwise post-hoc comparisons were not performed in this analysis. The patterns reported here therefore reflect the overall strength of taxonomic signal within each band and pH condition, rather than relationships between specific taxonomic groups.

SUBCELLULAR LOCALISATION

Localisation prediction provides a useful third feature to classify proteins, following paralogue and taxa. Subcellular localisation was predicted using DeepLoc2 (Thummuluri et al., 2022), a multi-label predictor that performs well across a range of eukaryotic compartments. DeepLoc2 achieves a **MCC** of approximately 0.90 for plastid localisation and 0.76 for mitochondrial localisation (Thummuluri et al., 2022). However, DeepLoc2 predicted no mitochondrial localisation for any sequences in this dataset, which conflicts with experimental evidence for mitochondrial targeting in AtSAL1 (Estavillo et al., 2011). The specialised tool MitoFates (Fukasawa et al., 2015) was therefore used to assess mitochondrial targeting. MitoFates specifically detects cleavable N-terminal presequences, achieving 79% precision and 80% recall on independent test data (Fukasawa et al., 2015). It should be noted that MitoFates only detects N-terminal

presequences, which account for approximately 70% of mitochondrial proteins; proteins using internal targeting signals will not be identified. Despite these limitations, any statistically significant clustering by predicted localisation would be noteworthy, as it would suggest that subcellular localisation information is encoded in the electrostatic field of the active site.

Analysing this data presents a statistical challenge. Because enzymes can be predicted to localise to multiple compartments, treating each compartment prediction independently would count the same enzyme multiple times. This means proportions across compartments would not sum to one per enzyme, violating standard statistical assumptions. To address this, I implemented one-hot encoding, which assigns a unique code to each combination of localisations observed in the dataset. This produced 12 unique localisation categories. This approach avoids double-counting but is conservative in that it cannot distinguish cases where one localisation target within a multi-compartment prediction is correct and others are not. For instance, consider enzyme A predicted to localise to the plastid and cytosol, and enzyme B predicted to localise to the plastid and mitochondrion. If the true localisation of both enzymes is solely the plastid, both predictions are classified as incorrect under this scheme. Moreover, these enzymes are not grouped together despite sharing the same true localisation. Any approach involve trade-offs, so I chose the more conservative option.

PERMANOVA analysis ($n = 506$; Table 5.8) showed that predicted localisation contributes to explaining variance across all bands and pH conditions. Bands 3 and 4 performed best and comparably, with R^2 values ranging from 0.120 to 0.124. Bands 1 and 2 showed weaker clustering power ($R^2 = 0.034$ -0.058 and 0.042-0.047 respectively), consistent with the pattern observed in the taxa analysis. The mid-frequency bands (3-6) showed remarkable stability across pH conditions, with R^2 varying by only 0.001-0.003 within each band. Band 1 showed modest pH dependence, with R^2 increasing from 0.034 at pH 7.0 to 0.058 at pH 7.9, before decreasing slightly to 0.049 at pH 8.2.

The similarity network in Figure 5.23, coloured by predicted localisation, reveals distinct clustering by paralogue. CNP sequences predominantly localise to the plastid, with a subset predicted as cytoplasmic. These cytoplasmic-predicted CNP electrostatic maps show poor clustering, suggesting greater heterogeneity among this subset. SAL functions as a plastid retrograde signalling enzyme (Chan et al., 2016), yet only 18.1% of SAL sequences were predicted to localise to the plastid (Table 5.9). This low proportion is unexpected given the enzyme's reported function. Interpreting these predictions requires caution. DeepLoc2 achieves a MCC of approximately 0.90 for plastid localisation

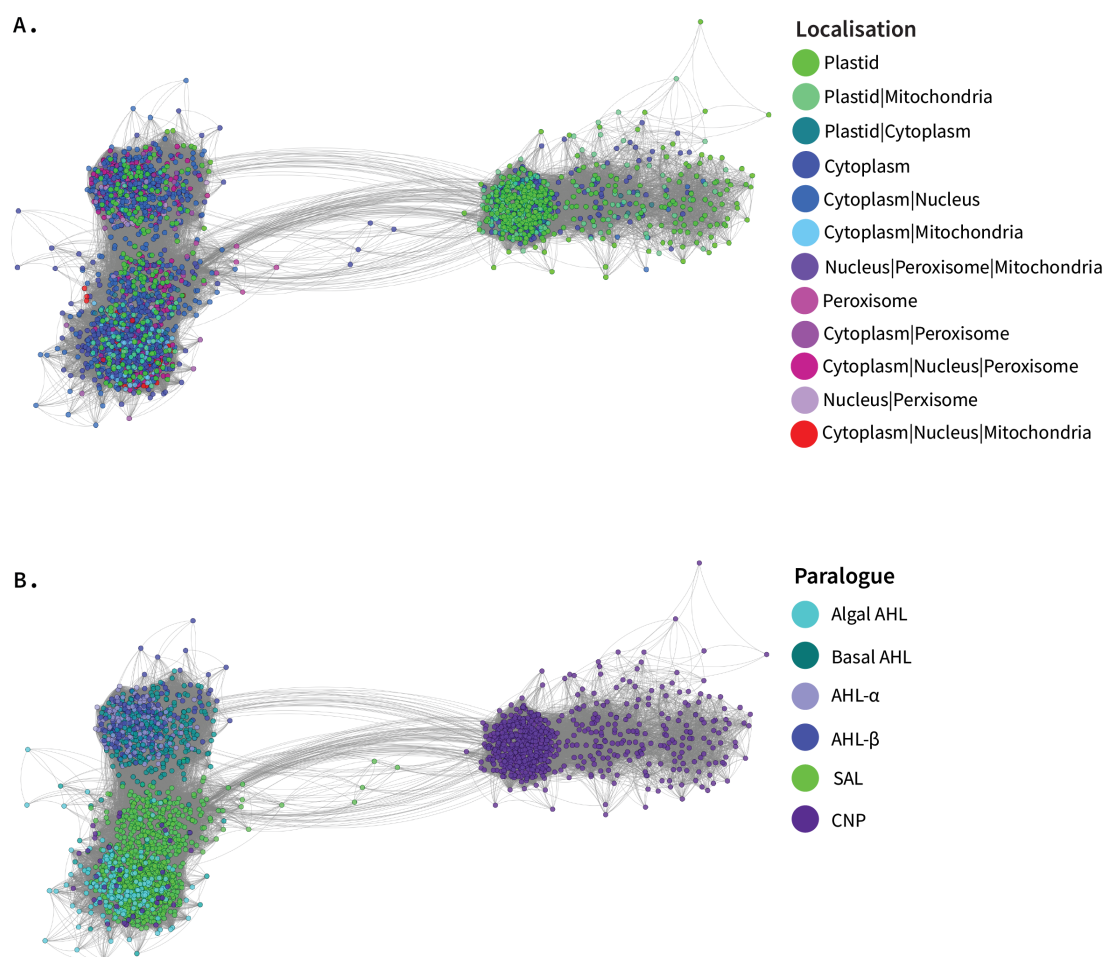


Figure 5.23: Fourier decomposed, frequency band 4 (scale 3.7–5.0 Å) active site electrostatic similarity network with Sinkhorn distances. Nodes represent frequency band 4 of decomposed active site electrostatic maps, with 2,530 nodes total of all 506 enzymes at pH 7.0, 7.3, 7.6, 7.9, and 8.2. Visualised in Cytoscape using Prefuse Force Directed Layout with edges filtered by the 0.08 quantile, coloured by localisation.

Table 5.8: PERMANOVA R^2 values by band and pH condition for localisation classification (n = 506).

Frequency band	pH7.0	pH7.3	pH7.6	pH7.9	pH8.2
Band 1	0.034*	0.040**	0.051**	0.058***	0.049**
Band 2	0.042***	0.044***	0.047***	0.042***	0.042***
Band 3	0.121***	0.121***	0.122***	0.121***	0.123***
Band 4	0.120***	0.122***	0.124***	0.124***	0.123***
Band 5	0.111***	0.110***	0.111***	0.110***	0.109***
Band 6	0.103***	0.102***	0.103***	0.104***	0.104***

Significance levels: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. FDR correction applied within each table independently.

(Thumuluri et al., 2022), indicating strong overall discriminative performance. However, MCC is a population-level metric and does not directly translate to per-prediction confidence. In particular, the false negative rate (plastid proteins incorrectly predicted as non-plastid) is not explicitly reported, so the absence of a plastid prediction cannot be interpreted with the same confidence as a positive prediction. CNP shows 69.2% plastid localisation, which may be consistent with fully conserved plastid targeting if false negatives account for the remainder, though this interpretation remains speculative without knowing the true false negative rate.

Table 5.9: Proportion of plastid localisation by paralogue (n = 506).

Paralogue	Proportion	n
AHL- α	0.000	43
AHL- β	0.209	43
Algal AHL	0.111	36
Basal AHL	0.231	65
CNP	0.692	159
SAL	0.181	160

COMBINED FEATURES

As clustering is multivariate and three labelled features are available, PERMANOVA analysis can identify how much of the variance in the dataset is explained by the unique combinations of paralogue, taxa, and localisation. This approach creates a compound grouping variable for each unique combination of the three features and tests whether these combinations cluster distinctly. It does not partition variance into separate contributions from each feature, but rather quantifies how well the combined classification explains the observed electrostatic differences.

PERMANOVA analysis (Table 5.10) showed that the combined feature classification explained up to 52.9% of the variance in the dataset, with the highest R^2 values observed in Band 4 at pH 7.9 and 8.2. The analysis tested 333 unique

combinations of paralogue, taxa, and localisation, with groups containing fewer than three members excluded. Bands 3 and 4 performed best and comparably ($R^2 = 0.518$ - 0.529), followed by Band 5 ($R^2 = 0.477$ - 0.481), then Band 6 ($R^2 = 0.462$ - 0.471). Bands 1 and 2 continued to show weaker clustering power ($R^2 = 0.214$ - 0.285 and 0.223 - 0.256 respectively), consistent with all previous analyses. The mid-frequency bands (3–6) showed remarkable stability across pH conditions, with R^2 varying by less than 0.01 within each band. Band 1 showed the same pH dependence pattern observed in previous analyses, with R^2 increasing from 0.214 at pH 7.0 to 0.285 at pH 7.9, before decreasing to 0.257 at pH 8.2.

Table 5.10: PERMANOVA R^2 values by band and pH condition for combined paralogue, taxa and localisation classification (n = 506).

Frequency band	pH7.0	pH7.3	pH7.6	pH7.9	pH8.2
Band 1	0.214**	0.225***	0.253***	0.285***	0.257***
Band 2	0.249***	0.252***	0.250***	0.256***	0.223***
Band 3	0.518***	0.519***	0.524***	0.524***	0.526***
Band 4	0.519***	0.523***	0.528***	0.529***	0.529***
Band 5	0.477***	0.477***	0.481***	0.481***	0.479***
Band 6	0.462***	0.463***	0.471***	0.471***	0.471***

Significance levels: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. FDR correction applied within each table independently.

Comparing across analyses, paralogue was the strongest individual predictor of electrostatic similarity (R^2 up to 0.335; Table 5.3), followed by taxa (R^2 up to 0.146; Table 5.4), then localisation (R^2 up to 0.124; Table 5.8). The combined analysis explained substantially more variance than any individual feature, indicating that all three contribute information about electrostatic field structure. The combined R^2 of 0.529 is less than the sum of the individual maximum R^2 values ($0.335 + 0.146 + 0.124 = 0.605$). This is expected because the three features share hierarchical structure and overlapping variance, with taxa nested within paralogues and localisation patterns differing between paralogues.

pH SENSITIVITY

The preceding analyses revealed that some features showed pH-dependent changes in clustering power. In particular, AHL and SAL showed increasing R^2 values for taxa clustering at higher pH in Band 1, while CNP remained stable across all pH conditions (Tables 5.5, 5.6, 5.7). To directly quantify pH sensitivity, I tested whether pH itself could serve as a clustering feature within each paralogue. This analysis asks whether electrostatic maps from the same protein at different pH values cluster together by pH condition.

PERMANOVA analysis using pH as the grouping feature (Table 5.11) showed that pH explains only a small proportion of variance in electrostatic similarity across all paralogues ($R^2 = 0.001\text{--}0.032$). With sample sizes of 158–187 proteins each represented at five pH conditions, even small effects can reach statistical significance. The significant p-values reported here therefore reflect detectable effects rather than strong clustering, and the low R^2 values indicate that pH alone is a weak predictor of electrostatic similarity compared to paralogue, taxa, or localisation.

Table 5.11: PERMANOVA R^2 values by band and paralogue for pH classification (n = 506).

Frequency band	SAL (n=161)	CNP (n=158)	AHL (n=187)
Band 1	0.019***	0.014**	0.032***
Band 2	0.013***	0.011**	0.007*
Band 3	0.021***	0.007	0.013***
Band 4	0.012***	0.003	0.007**
Band 5	0.006	0.003	0.005
Band 6	0.001	0.003	0.002

Significance levels: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. FDR correction applied within each table independently.

Degrees of freedom: SAL ($df_{\text{between}} = 4$, $df_{\text{within}} = 800$), CNP ($df_{\text{between}} = 4$, $df_{\text{within}} = 785$), AHL ($df_{\text{between}} = 4$, $df_{\text{within}} = 930$).

Despite the low effect sizes, the patterns differed between paralogues. AHL (n = 187) showed the strongest pH clustering in Band 1 ($R^2 = 0.032$, $p < 0.001$), with weaker but significant effects in Bands 2–4. SAL (n = 161) showed the most consistent pH clustering across bands, with significant effects in Bands 1–4 and the highest R^2 value in Band 3 ($R^2 = 0.021$). CNP (n = 158) showed the weakest pH clustering overall, with significant effects limited to Bands 1–2 and no significant clustering in Bands 3–6. These differences in pH sensitivity align with the pH-dependent patterns observed in the taxa analysis, where AHL and SAL showed variable clustering power across pH conditions while CNP remained stable.

Across all paralogues, Band 1 showed the strongest pH clustering, while Bands 5 and 6 showed minimal pH sensitivity with no values reaching statistical significance. This concentration of pH effects in the lower frequency bands suggests that pH-induced changes to active site electrostatics primarily affect larger-scale electrostatic features.

To explore whether pH sensitivity varied within AHL, I repeated the analysis for the four AHL sub-clades (Table 5.12). The sub-clades showed markedly different pH sensitivity profiles. AHL- α (n = 43) showed the strongest pH clustering of any group analysed, with Band 1 $R^2 = 0.078$ ($p < 0.001$), substantially higher than the AHL paralogue as a whole ($R^2 = 0.032$), and a second-

Table 5.12: PERMANOVA R^2 values by band and AHL sub-clade for pH classification (n = 506).

Frequency band	AHL- α (n=43)	AHL- β (n=43)	Basal AHL (n=65)	Algal AHL (n=36)
Band 1	0.078***	0.023	0.057***	0.040
Band 2	0.013	0.002	0.013	0.052**
Band 3	0.038**	0.007	0.018	0.059***
Band 4	0.022	0.003	0.010	0.035*
Band 5	0.010	0.004	0.006	0.024
Band 6	0.007	0.005	0.002	0.005

Significance levels: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. FDR correction applied within each table independently. Degrees of freedom: AHL- α ($df_{between} = 4$, $df_{within} = 210$), AHL- β ($df_{between} = 4$, $df_{within} = 210$), Basal AHL ($df_{between} = 4$, $df_{within} = 320$), Algal AHL ($df_{between} = 4$, $df_{within} = 175$).

ary effect in Band 3 ($R^2 = 0.038$, $p < 0.01$). Basal AHL (n = 65) also showed strong pH clustering in Band 1 ($R^2 = 0.057$, $p < 0.001$), though effects in other bands did not reach significance. In contrast, Algal AHL (n = 36) showed a distinct pattern, with its strongest pH effects in the mid-frequency bands (Bands 2–4) rather than Band 1, and was the only sub-clade to show significant pH clustering in Band 4. AHL- β (n = 43) showed no significant pH clustering in any band. This divergence indicates that the moderate pH sensitivity observed in the combined AHL analysis reflects contributions from multiple sub-clades with distinct patterns: AHL- α and Basal AHL driving the Band 1 effect, and Algal AHL contributing mid-frequency sensitivity, while AHL- β contributes little to no pH-dependent variation.

5.4 Discussion

This chapter developed a novel methodology for classifying homologous enzymes based on the electrostatic properties of their active sites. Through application to plant nucleotide phosphatases, I demonstrated that electrostatic fields encode biologically meaningful information that builds upon traditional sequence-based and structure-based approaches. The Fourier decomposition method successfully discriminated between SAL, AHL, and CNP paralogues, revealed unexpected evolutionary relationships between enzyme sub-clades, and detected pH-dependent clustering patterns that would be invisible to sequence or structural comparison alone.

5.4.1 Summary of key findings

The central finding of this chapter is that Fourier decomposition of electrostatic fields, combined with optimal transport comparison, provides a robust method

for classifying homologous enzymes. The optimal frequency band (Band 4, corresponding to spatial wavelengths of 3.7–5.0 Å) explained up to 33.5% of variance in electrostatic distance by paralogue identity. This represents a substantial improvement over direct comparison of three-dimensional electrostatic maps (Approach 2), which failed to discriminate paralogues due to noise amplification, a single net charge comparison, and sampling bias during point extraction.

The key methodological insight was that separating electrostatic features by spatial scale allows targeted analysis of the frequency components that carry the strongest biological signal. Rather than computing a single distance metric that mixes all spatial scales together, the Fourier approach isolates features at different wavelengths. This separation revealed that mid-frequency components (3.7–6.7 Å), associated with variations in secondary structure, and carries the strongest paralogue and taxonomic signal. The low-frequency components associated the overall active site are dominated by noise/artefacts likely from the windowing technique, and correlate with pH discrimination, particularly in AHL.

A combined statistical model incorporating paralogue identity, taxonomic origin, and predicted subcellular localisation explained approximately 53% of variance in electrostatic similarity. Noise robustness analysis, conducted by comparing replicate calculations with different random seeds, indicated that approximately 20% of variance is attributable to methodological noise from the computational pipeline and error in the underlying structural models. This leaves approximately 27% as unexplained biological variation. I interpret this residual variance as representing unmeasured functional differences between enzymes, such as catalytic parameters, substrate specificity, or regulatory properties that we do not have labelled in this dataset to confirm.

5.4.2 Methodological progression from sequence to electrostatic classification

The development of the Fourier decomposition approach is the culmination of methodological refinement across this thesis. It built on successful techniques from earlier approaches while addressing their limitations.

Structural modelling: capabilities and constraints

Chapter 3 demonstrated the utility of AF2 structural modelling for understanding enzyme dysfunction. The analysis of PSY variants successfully identified

mechanistic explanations for reduced activity through steric effects (M266I), fold stability disruption (A352T), and catalytic geometry impairment (P178S). These predictions were subsequently validated through experimental characterisation, establishing that computational structural analysis can generate accurate functional hypotheses.

However, Chapter 3 also revealed the limits of static structural models. The psy-4 (M266I) and psy-90 (A352T) variants showed reduced interaction with OR despite no obvious structural impediment at the predicted binding interface. This disconnect between structural prediction and functional outcome suggested that steric effects from the structural model were insufficient to explain the functional impact alone. In particular, structural models are unable to account for electrostatic field distributions that govern protein-protein interactions. This observation motivated the development of methods to compare electrostatic properties systematically.

Sequence similarity networks for classification

Chapters 3 and 4 successfully employed SSN, Evolocity, and phylogenetic analysis to classify protein paralogues and highlight evolutionary relationships. The DXO1 analysis (Chapter 3) identified angiosperm-specific innovations through the serine-rich N-terminal domain. The nucleotide phosphatase analysis (Chapter 4) used SSNs to distinguish SAL, AHL, and CNP as distinct paralogues, while phylogenetics provide increased discrimination between groups of enzymes and predicted neo-functionalisation following gene duplication events. CNP was found to be a novel, uncharacterised paralogue of SAL and AHL. Gene Ontology (Table 4.4) and tissue expression (Figure 4.7) suggested possible roles for CNP and reduced the experimental space needed to be explored in order to characterise this enzyme.

Yet these sequence-based methods – SSNs, Evolocity, and phylogenetics – share a fundamental limitation in that they operate at the sequence-structure level of abstraction. As noted in Chapter 4, sequence similarity provides an imperfect proxy for functional similarity. Two enzymes with 80% sequence identity might have identical catalytic properties or might differ substantially, depending on where the differing residues are located and how they affect the active site environment. The phylogenetic approach could identify that AHL- α and AHL- β arose from gene duplication and statistical differences were evident in sequence properties (Table 4.1), but could not determine whether these sequence differences translated to catalytic divergence or merely reflected neutral drift with functional conservation.

Chapter 4 explicitly identified this gap and proposed the solution: "Catalytic functional divergence is defined by the active site and can be approximated as the shape of the binding pocket and then electrostatic environment. By comparing these, in particular the active site electrostatic maps... we could further refine these functional classifications." The methodology developed in this chapter directly addresses this proposal.

Why direct electrostatic comparison failed

The progression from sequence-based to electrostatics-based classification required overcoming substantial methodological challenges. Two initial approaches failed before the successful Fourier decomposition method emerged.

The surface charge imaging approach (Approach 1) demonstrated that electrostatic information could distinguish protein families, validating the core hypothesis. However, projecting three-dimensional electrostatic fields onto two-dimensional surface representations discarded depth information essential for comparing active site cavities. The problem is analogous to looking into a cave from outside, where the view reveals the entrance clearly but obscures any twists, turns, or features deeper within the cavity. For enzyme active sites, this means that surface-projected electrostatic potentials cannot capture obscured parts of the active site, significantly degrading the ability to compare charges while losing spatial information aligned with the axis of view.

Direct optimal transport comparison of three-dimensional electrostatic maps (Approach 2) preserved spatial information but failed to discriminate paralogues. The fundamental problem was that optimal transport computes a single distance value mixing contributions from all spatial scales. Small local perturbations from structural modelling uncertainty, numerical artefacts, or sampling effects contributed to the distance metric with the same weight as genuine functional differences. Biologically relevant information could not be extracted without a mechanism to separate signal from noise, and to separate charge distribution across spatial scales.

How Fourier decomposition resolves earlier limitations

The Fourier decomposition approach addresses the limitations of both earlier methods by separating electrostatic features by spatial scale before comparison. This separation provides two key advantages.

First, it enables noise isolation as analysis revealed that noise concentrates in specific frequency bands (Figure 5.20). The highest frequencies corresponding to spatial scales below 2 Å were not included in the analysis following initial

band optimisation as it was evident early on that error from AF2 modelling and numerical noise was being captured. While including high frequency bands should improve discrimination between active sites through capturing more subtle residue-specific features – which should correspond with specific catalytic features such as catalytic rate and substrate binding strength – the contribution of noise overrode any benefit from retaining this band. From the remaining bands noise or artefacts were particularly evident in the low-frequency bands affected by boundary effects from Hann windowing of the cropped active-sites. Mid-frequency bands (Bands 3–4, corresponding to 3.7–6.7 Å wavelengths) contain approximately 20% noise while carrying the strongest biological signal. This knowledge enables appropriate weighting and interpretation, with low-frequency results treated more cautiously than mid-frequency findings.

Secondly, separating spatial scales reveals which features drive functional classification. The observation that mid-frequency bands outperform others indicates that secondary structure-scale features, rather than overall charge distribution or individual residue positions, distinguish these paralogues. This insight has biological implications, suggesting that functional divergence following gene duplication apparently occurs through reorganisation at secondary structure scales rather than through gross architectural changes or point substitutions. Notably, in paralogues with apparent pH sensitivity, the low frequency bands provided the greatest discriminatory power. This suggests that pH sensitivity is most readily identifiable in protein or domain-wide charge features. Improvements in the windowing technique may provide greater clarity here. Given that different features (paralogue, tax, localisation and pH sensitivity) are best discriminated by different bands suggests that evolution is acting on these features at different scales. This raises the question whether allosteric sequence changes, that may have previously been written off as neutral drift, between orthologues from closely related species could in fact be adaptations to selective pressures that may not be evident due to limited experimental characterisation. This could include subcellular organelle pH variation.

Application to nucleotide phosphatase classification

The methodology developed here directly addresses the questions raised in Chapter 4 about nucleotide phosphatase functional divergence. The phylogenetic analysis predicted that AHL- α and AHL- β had undergone neo- or sub-functionalisation following the γ -polyploidy event, based on their retention as distinct clades and statistical differences in sequence properties. The electrostatic analysis provides a means to putatively test this.

The clustering of AHL- α , AHL- β , and Basal AHL in electrostatic similarity space indicates that these sub-clades have not diverged substantially in their active site electrostatic properties, yet the AHL- α , AHL- β duplication in Eudicots is almost universally conserved. Their similar mid-frequency signatures suggest comparable substrate binding environments. However, the divergent pH sensitivity patterns, with AHL- α showing strong pH-dependent clustering while AHL- β shows none (Table 5.12), reveals functional divergence invisible to sequence comparison. The sub-clades appear to have maintained similar catalytic machinery while diverging in how that machinery responds to environmental pH.

The duplicate retention following γ -polyploidy likely reflects tissue-specific or developmental expression patterns rather than catalytic neo-functionalisation. AHL- α and AHL- β may perform the same reaction in different cellular contexts, with AHL- α specialised for pH-dynamic environments such as guard cells while AHL- β operates in more pH-stable conditions. Notably tissue expression of AtAHL (AHL- α) is particularly high in the guard cells (Figure 4.7). The loss of AHL- β specifically in Brassicaceae (Chapter 4) may reflect lineage-specific changes in expression patterns that made spatial separation unnecessary. The lack of AHL- β in *A. thaliana* and the lack of experimental characterisation from other species leave this question open for now, but provides an obvious experiment to perform. This also highlights the risk of enzyme classification using model species when paralogues have not been mapped across the species of interest. *A. thaliana* is likely a very poor model for future studies on AHL.

Implications for wheat SAL functional divergence

The electrostatic methodology also offers a path to understanding the TaSAL1 versus TaSAL2 functional divergence characterised in Chapter 4. These wheat sequences were not included in the broader nucleotide phosphatase analysis, as the representative sequence set was derived programmatically to avoid taxonomic bias (Methods, Section 2.1.2; Section 2.1.3) and were not captured. However, the framework developed here provides a basis for targeted investigation. The field trial results demonstrated opposing phenotypic effects, with TaSAL2 deletion improving yield while TaSAL1 deletion reduced it. Phylogenetic and co-expression analyses indicated functional divergence following gene duplication, with TaSAL1 maintaining core metabolic functions and TaSAL2 evolving stress-responsive roles.

The mechanistic basis for this divergence remains unclear. Chapter 4 pro-

posed neo-functionalisation combined with gene dosage effects in polyploid wheat, but could not determine whether **TaSAL1** and **TaSAL2** differ in their catalytic properties or only in their expression patterns and regulatory contexts. Electrostatic comparison of **TaSAL1** and **TaSAL2** active sites would address this question directly.

If **TaSAL1** and **TaSAL2** show similar electrostatic signatures, analogous to the **AHL- α** /**AHL- β** pattern, this would support functional divergence through expression and regulatory changes rather than catalytic modification. The enzymes would perform the same reaction but in different contexts, explaining why deleting **TaSAL2** (stress-responsive) improves yield under field conditions while deletion of **TaSAL1** (core metabolism) impairs growth.

Alternatively, if **TaSAL1** and **TaSAL2** show distinct electrostatic signatures, this would indicate catalytic divergence with potential differences in substrate affinity, pH optimum, or catalytic features. Such divergence could explain the different phenotypic consequences of their deletion through effects on PAP metabolism and retrograde signalling that extend beyond simple expression pattern differences.

pH-resolved electrostatic analysis would be particularly interesting as sequence-based methods would not capture this effect. If **TaSAL2**'s stress-responsive function involves pH-dependent activity modulation, analogous to the **AHL- α** pattern, this would provide mechanistic insight into how wheat SAL paralogues integrate stress signals. The methodology developed in this chapter provides the tools to address these questions, representing a direct application of the Fourier decomposition approach to agriculturally relevant functional prediction.

5.4.3 Biological interpretation of clustering patterns: Functional continuity of PAP catabolism across Viridiplantae

The electrostatic analysis revealed several unexpected clustering patterns that provide biological insights beyond what sequence or structural analysis alone can offer.

The electrostatic analyses complicates a simpler assumption of SAL orthologues providing a direct evolutionary trajectory for **PAP** catabolism. Sequence-based phylogeny groups proteins by sequence similarity, but electrostatic comparison of active sites groups them by the physical properties that more directly determine catalytic function. These two approaches produced different results for some **AHL** and **CNP** clades. The different clustering of these groups in the electrostatic analysis provides an explanation for the functional gap identified

in Chapter 4.

The functional gap in PAP catabolism

PAP is a toxic byproduct of sulphur assimilation that alters RNA processing, so it must be degraded to maintain cellular function (Murguía et al., 1995; Chan et al., 2016). In land plants, PAP accumulation during stress serves as a retrograde signal coordinating nuclear gene expression with chloroplast status (Estavillo et al., 2011; Chan et al., 2016). The enzyme responsible for PAP degradation must therefore be tightly regulated.

In the plant lineage going back to a common ancestor with yeast, we have two enzymes that have experimentally demonstrated efficient PAP phosphatase activity. ScHAL2 from *Saccharomyces cerevisiae* (*S. cerevisiae*) shows a K_M of approximately 0.3–2.5 μM for PAP, with variation between studies likely reflecting different assay conditions (Murguía et al., 1995; Yang et al., 2014). AtSAL1 from *A. thaliana* shows comparable kinetics with a K_M of 2–10 μM (Quintero et al., 1996; Gil-Mascarell et al., 1999; Chan et al., 2016). Both enzymes share close sequence similarity which lead to Quintero et al. (1996) identifying AtSAL1 through homology with ScHAL2, and the SSN analysis in Chapter 4 confirmed their close relationship (Figure 4.1).

This sequence similarity between ScHAL2 and SAL, combined with their shared catalytic function, suggests a single evolutionary origin for PAP catabolism. The alternative hypothesis is that PAP phosphatase activity arose independently in fungi and plants through convergent evolution from other phosphatases. This alternative cannot be excluded, but parsimony favours shared ancestry. Phosphatases are ancient and diverse enzymes, and the specific combination of sequence similarity and shared substrate specificity between ScHAL2 and SAL is unlikely to have arisen through convergent evolution. For this discussion, I argue under the working hypothesis that PAP catabolism shares a common origin in fungi and plants, acknowledging that future, more detailed phylogenetics when more sequences become available, may overturn this.

Under this hypothesis, a continuous lineage of organisms capable of degrading PAP should exist from the common ancestor of fungi and plants through to modern land plants. The phylogenetic distribution of known PAP-catabolising enzymes creates an apparent problem with this expectation. SAL is present only in streptophytes and absent from chlorophyte algae. Chromista, Glaucophyta, and Rhodophyta lack SAL entirely, possessing only sequences that comprise of the ancient algal CNP clade. Chlorophytes possess both CNP and Algal AHL, but not SAL. If only SAL and ScHAL2 can effectively degrade PAP, these

lineages would lack PAP catabolic capacity entirely. This represents a functional gap spanning hundreds of millions of years of eukaryotic evolution (Figure 5.24A).

Several scenarios could explain this gap. PAP catabolism may have been lost in these lineages, with organisms either tolerating PAP accumulation or employing alternative degradation pathways not yet characterised. Alternatively, the phylogenetic reconstruction may be incorrect, with SAL actually present in these lineages but undetected due to sequence divergence or incomplete sampling. A third possibility is that enzymes currently classified outside SAL based on sequence similarity nevertheless possess PAP catabolic function. The electrostatic analyses presented in this chapter suggest this last scenario.

Electrostatic evidence for functional relationships

Throughout the analyses presented in this chapter, Algal AHL and the ancient algal CNP clade consistently clustered with SAL in electrostatic similarity rather than with their sequence-defined paralogues.

Algal AHL sequences clustered within the SAL cluster in the electrostatic similarity network (Figure 5.17), despite grouping near other AHL sequences in both sequence similarity and structural similarity networks (Figure 5.18). This pattern was most pronounced in frequency bands 3 and 4, corresponding to spatial wavelengths of 3.7–6.7 Å. These mid-frequency bands capture electrostatic features at the spatial scales of secondary structure and binding pockets which are expected to govern substrate recognition and catalytic mechanism.

Ancient CNP sequences from Chromista, Rhodophyta, and Glaucophyta showed a similar pattern. In the sequence and structural similarity networks, these sequences clustered as outliers not overlapping SAL (Figure 5.19), while the phylogeny (Figure 4.3) classified them CNP members. In the electrostatic similarity network, however, they positioned within the SAL cluster rather than with the main CNP group as expected from the phylogenetic analysis. The same proteins showed anomalous electrostatic clustering consistently across all three comparison methods, confirming this represents a genuine property of these sequences rather than a methodological artefact.

Available kinetic data can aid in interpreting these results but definitive support is not possible as direct comparison between studies is challenging given that kinetic parameters were determined under different assay conditions including varying pH, temperature, ionic strength, and substrate purity. ScHAL2 exhibits a K_M reported variously as 0.3 μM (Yang et al., 2014) and 2.5 μM (Murguía et al., 1995), with k_{cat} of 11 s^{-1} (Yang et al., 2014). AtSAL1 shows a K_M

of 2–10 μM and k_{cat} of 2.05 s^{-1} (Quintero et al., 1996; Gil-Mascarell et al., 1999; Chan et al., 2016). The approximately 10-fold variation in reported ScHAL2 K_{M} values illustrates the difficulty of comparing kinetic parameters across studies. Nevertheless, both ScHAL2 and AtSAL1 show K_{M} values in the low micromolar range, consistent with physiologically relevant PAP affinity, and both enzymes cluster together in electrostatic similarity networks.

AtAHL has a reported K_{M} of approximately 160 μM for PAP (Gil-Mascarell et al., 1999), indicating substantially weaker substrate affinity than either SAL or ScHAL2—approximately 15- to 500-fold higher depending on which ScHAL2 value is used for comparison. Full Michaelis-Menten kinetics including k_{cat} and V_{max} have not been reported for AtAHL, preventing catalytic efficiency ($k_{\text{cat}}/K_{\text{M}}$) determination. The high K_{M} alone, however, suggests that AtAHL would require supraphysiological PAP concentrations for substantial activity.

The electrostatic clustering of Algal AHL with SAL rather than with land plant AHL sub-clades generates the testable hypothesis that Algal AHL retains PAP affinity comparable to SAL, having not undergone the functional shift that characterises land plant AHL. If correct, this would resolve the apparent functional gap in PAP catabolism among chlorophytes as Algal AHL would fulfil the PAP degradation role that SAL performs in streptophytes. Kinetic characterisation of representative Algal AHL and Ancient CNP, including determination of K_{M} , k_{cat} , and V_{max} for PAP under standardised assay conditions with a AtSAL1 control, would directly test this prediction.

A model for functional continuity

The electrostatic evidence suggests a model for the evolutionary continuity of PAP catabolism that resolves the apparent functional gap (Figure 5.24B–C). Rather than requiring loss and re-evolution of PAP catabolic capacity, the enzymes performing this function in different lineages may be classified into different sequence-based groups while retaining similar catalytic properties.

Under this model, ancient CNP sequences in Chromista, Glaucophyta, and Rhodophyta possess PAP catabolic function despite their sequence classification. Their electrostatic similarity to SAL reflects conservation of the active site properties required for PAP binding. Similarly, Algal AHL in chlorophytes functions as the PAP phosphatase in these organisms, with electrostatic properties matching those of streptophyte SAL. The emergence of SAL as a distinct sequence group in streptophytes may represent continued evolution of the PAP-catabolising lineage, with sequence divergence from Algal AHL obscuring the functional continuity. This model is shown diagrammatically in this section in

Figure 5.17.

This model proposes that **PAP** catabolism has been maintained continuously from **ScHAL2** in fungi through Ancient **CNP** in early-diverging eukaryotic algae, Algal **AHL** in chlorophytes, and **SAL** in streptophytes (Figure 5.24C). The apparent functional gap arises only when sequence-based classification is assumed to reflect function. Electrostatic comparison of active sites reveals functional relationships that sequence comparison obscures.

The model also accounts for the divergent properties of land plant **AHL**. **AtAHL** shows poor affinity for **PAP**, with a K_M of approximately 160 μM (Gil-Mascarell et al., 1999). This is roughly two orders of magnitude higher than **AtSAL1** or **ScHAL2**. In the electrostatic similarity networks, land plant **AHL** sub-clades (Basal **AHL**, **AHL- α** , **AHL- β**) overlap substantially with each other while remaining separate from both **SAL** and Algal **AHL** (Figure 5.17B). This overlapping electrostatic signature among all three land plant **AHL** sub-clades suggests they share the same or highly similar catalytic function (Figure 5.24B), despite their sequence-level divergence following the γ -polyploidy event. The poor **PAP** affinity of **AtAHL** likely reflects this shared land plant **AHL** function, which differs from the **PAP** catabolism performed by Algal **AHL** and **SAL**. Guard cell localisation of **AtAHL** and its drought-responsive expression pattern, discussed in Chapter 4, support neo-functionalisation toward roles in stomatal regulation rather than **PAP** homeostasis and retrograde signalling.

Integration with phylogenetic uncertainty

This functional model connects directly to the phylogenetic uncertainty discussed in Chapter 4. The poor branch support at the **SAL**–**AHL** divergence node left unresolved whether chlorophytes lost **SAL** after an early duplication (Figure 4.9A) or whether the duplication producing **SAL** occurred within streptophytes after their divergence from chlorophytes (Figure 4.9B). Sequence-based phylogenetic analysis alone could not distinguish these scenarios due to inadequate chlorophyte sampling and limited phylogenetic signal at this node.

The electrostatic analysis provides an independent line of evidence. The clustering of Algal **AHL** with **SAL** in electrostatic space, rather than with land plant **AHL**, is more consistent with the second scenario. Under this interpretation, Algal **AHL** represents the ancestral **PAP**-catabolising enzyme that gave rise to both **SAL** and land plant **AHL** following streptophyte emergence. **SAL** retained the ancestral **PAP** catabolic function while land plant **AHL** diverged toward new roles. Chlorophytes never possessed **SAL** because the relevant duplication had not yet occurred.

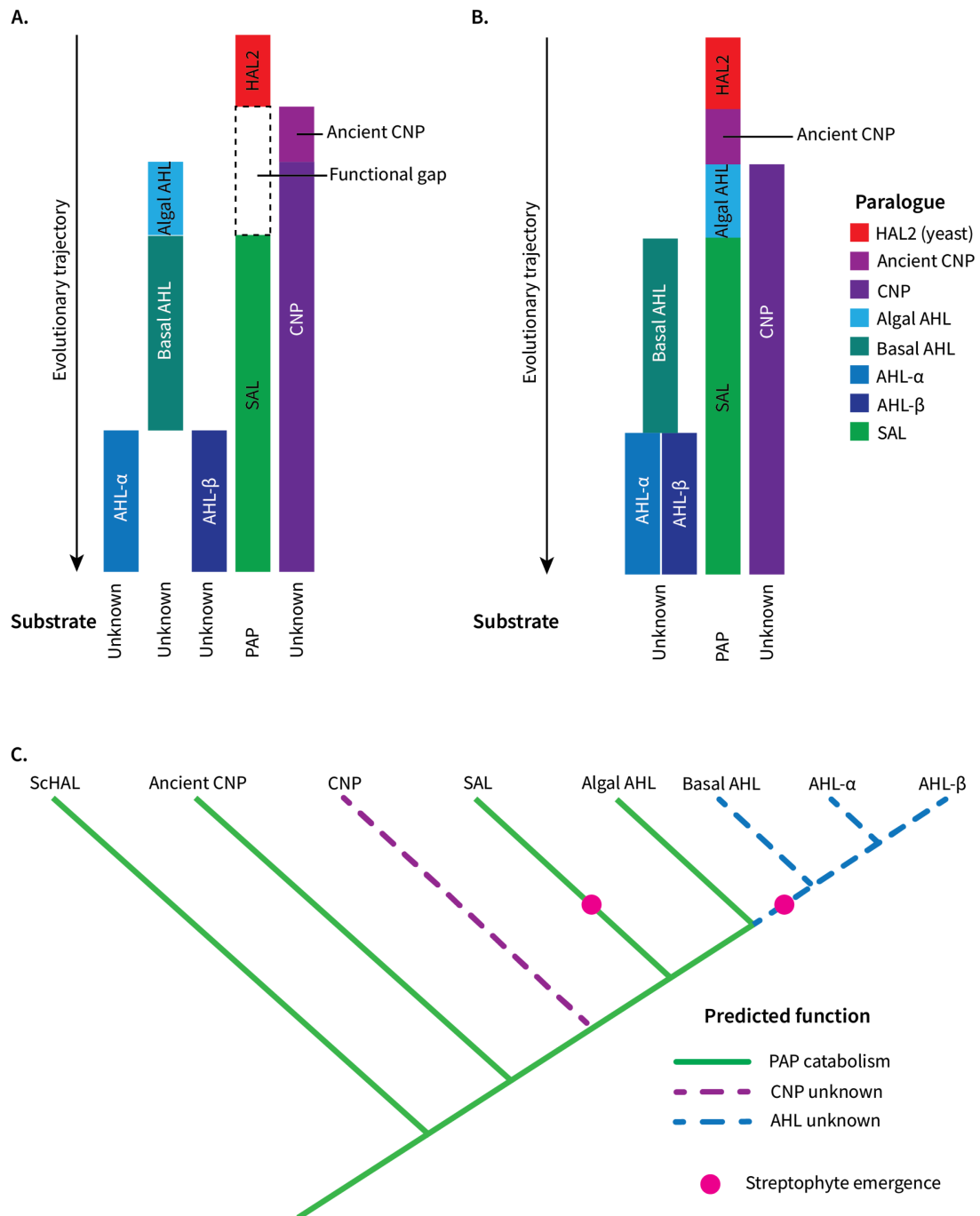


Figure 5.24: Functional evolution of PAP catabolism in plant nucleotide phosphatases. Vertical bars represent paralogues along an evolutionary trajectory (y-axis) with substrate specificity (x-axis). (A) The functional gap scenario. If only SAL and ScHAL2 possess efficient PAP catabolism, a gap exists between ScHAL2 (dashed box) and SAL. Ancient CNP and Algal AHL would lack PAP catabolic capacity. (B) Model based on electrostatic evidence. Positioning reflects functional relationships from electrostatic similarity. ScHAL2 and Ancient CNP are adjacent, indicating functional continuity. Algal AHL is positioned near SAL, reflecting electrostatic clustering and shared PAP function. Basal AHL, AHL- α , and AHL- β are grouped, reflecting overlapping electrostatic signatures and similar (unknown) function distinct from PAP catabolism. (C) Cladogram of proposed functional relationships mapped onto phylogeny. Green solid branches indicate predicted PAP catabolic function from ScScHAL2 through Ancient CNP, Algal AHL, to SAL. Purple dashed branches indicate CNP with unknown function. Blue dashed branches indicate land plant AHL lineages with unknown function. Streptophyte emergence (pink circle).

This interpretation does not definitively resolve the phylogenetic topology. Sequence similarity remains what it is regardless of functional properties, and the poor branch support at the SAL–AHL node reflects genuine uncertainty in the sequence data. However, the electrostatic evidence demonstrates that Algal AHL shares the active site properties most relevant to PAP catabolism with SAL rather than with land plant AHLs. When sequence-based and function-based classifications disagree, the functional evidence may better predict catalytic behaviour, but both lines of evidence contribute to understanding evolutionary history.

Testable predictions

This functional continuity model has testable predictions for biochemical characterisation. If Algal AHL performs PAP catabolism in chlorophytes, representative enzymes from this group should show kinetic parameters comparable to those of SAL and ScHAL2. Specifically, K_M values for PAP should fall in the low micromolar range (2–10 μM) rather than the high value reported for AtAHL (160 μM). Similarly, Ancient CNP sequences from Chromista, Glaucophyta, or Rhodophyta should show efficient PAP phosphatase activity if they function as the PAP catabolic enzyme in these lineages.

The model also predicts that the main CNP clade, which clusters separately from SAL in electrostatic space, should show reduced or absent PAP phosphatase activity. The domain insertion identified in Chapter 4 (Figure 4.5) may contribute to altered substrate specificity by increasing available sequence space. Testing CNP catalytic kinetics would therefore provide a negative control for the hypothesis that electrostatic similarity to SAL predicts PAP catabolic function.

Experimental validation would require heterologous expression and purification of representative enzymes from each group, followed by catalytic kinetic characterisation against PAP. A minimal test would include representatives each from Algal AHL, Ancient CNP, main CNP, and land plant AHL (beyond AtAHL, which has already been characterised, noting that Chapter 3 suggests it is an evolutionary outlier given the apparent absence of AHL- β in Brassicaceae). Comparison of these kinetic parameters with those of AtSAL1 would test whether electrostatic clustering predicts catalytic similarity. Given the diversity of organisms involved, some sequences may prove difficult to express in standard heterologous systems, but even partial sampling across the proposed functional groups would provide a meaningful test of the model. Indeed, more detailed analysis of pH sensitivity as demonstrated in section 5.3.3,

of candidate enzymes may help determine relevant physiological pH conditions that could impact not only enzyme kinetics but also indicate optimal pH for protein stability during purification.

If the predictions are not supported, alternative interpretations would be required. PAP catabolism may indeed have arisen multiple times independently, with the electrostatic similarity between Algal AHL, Ancient CNP, and SAL reflecting structural constraints on this enzyme fold rather than shared catalytic function. Alternatively, electrostatic similarity may not reliably predict kinetic parameters for this enzyme family, limiting the utility of this approach for functional classification. Either outcome would be informative for understanding both nucleotide phosphatase evolution and the broader applicability of electrostatic comparison methods.

Implications for enzyme classification

The finding that electrostatic similarity can reveal functional relationships obscured by sequence divergence has broader implications for enzyme classification. Sequence-based methods remain essential for establishing evolutionary relationships and identifying homologues, but they may not reliably predict catalytic function when active site properties can be conserved despite overall sequence divergence.

This principle is evident in the plant nucleotide phosphatases where homologues classified into different paralogous groups based on sequence (specifically Algal AHL versus SAL) share active site electrostatic properties that may determine their catalytic behaviour. Conversely, enzymes classified within the same sequence-based group (specifically Algal AHL versus land plant AHL) show divergent electrostatic properties consistent with different catalytic capabilities. Neither sequence nor electrostatic comparison alone provides complete information. Sequence establishes evolutionary relationships and identifies candidate homologues. The electrostatic analysis provides evidence about the physical properties most directly relevant to catalysis. Integration of both approaches enables more refined classification than either alone.

This aligns with the broader argument developed throughout this thesis where the sequence-structure-function paradigm treats sequence as the primary determinant of function through its determination of structure. However, the electrostatic field of the active site more directly determines catalytic properties. Sequence changes that do not alter active site electrostatics may have minimal functional consequences. Changes that do alter the electrostatic environment can affect catalysis even if overall sequence similarity remains high. By compar-

ing active site electrostatics directly, we access a level of functional information that sequence comparison cannot provide.

For the plant nucleotide phosphatases specifically, the electrostatic analysis suggests that functional classification should not assume correspondence with sequence-based paralogue classification. Algal **AHL** and Ancient **CNP** should be considered candidate **PAP** phosphatases despite their phylogenetic classification outside SAL. Land plant **AHL**, despite sharing a paralogue name with Algal **AHL**, appears to have diverged functionally and should not be assumed to share **PAP** catabolic properties with its algal relatives. These distinctions have practical implications for interpreting studies that manipulate nucleotide phosphatases, as discussed in Chapter 4 for wheat SAL. Understanding which enzymes share catalytic function, rather than merely sequence similarity, is essential for understanding metabolic networks through to predicting phenotypic outcomes of genetic manipulation.

Subcellular localisation and electrostatic properties

The significant electrostatic-derived clustering by predicted subcellular localisation might initially seem surprising given transit peptides, the N-terminal signals that direct proteins to plastids and other organelles, are cleaved after import and therefore do not contribute to the mature enzyme's electrostatic field. But enzymes evolve under selective pressure to function optimally in their destination compartment. For example, the plastid stroma has different properties from the cytosol. Stromal pH fluctuates diurnally as photosynthetic activity varies, ionic strength differs, and the complement of interaction partners is distinct. Cytosolic enzymes experience different conditions. These environmental differences exert selective pressure on enzyme properties including the electrostatic field of the active site.

For example, a plastid(stroma)-localised enzyme experiences pH values ranging from approximately 7 in the dark to above 8 during active photosynthesis (Trinh and Masuda, 2022; Heldt et al., 1973). The enzyme must function across this range, which constrains the pKa values of titratable residues in the active site. Alternatively this may provide a mechanism for photosynthesis-linked post-translational pH dependent regulation. Cytosolic enzymes experience more stable pH around 7.0–7.4 (Shen et al., 2013) and can be optimised for that specific value. Over evolutionary time, these different constraints shape different electrostatic signatures, even though the transit peptide responsible for the localisation itself is removed and never contributes to the active site electrostatic field.

AtSAL1 was shown to be dual plastid-mitochondria localised (Estavillo et al., 2011), however paralogue-wide predicted localisation shows that a high proportion of SAL enzymes were predicted not to be plastid-localised and deserves particular attention. The discrepancy between experimental data and computational prediction may indicate non-conserved localisation of SAL1 (with consequences on extrapolating its role as a drought sensor to other species), SAL-specific limitations in predictive accuracy of the prediction tool, or systematic annotation errors in the sequence databases.

Plastid localisation in higher plants depends on N-terminal transit peptides with characteristic sequence features that computational tools can recognise. However, this recognition depends on correct identification of the translation start site. If a sequence is annotated with an incorrect start codon downstream of the true start, the transit peptide appears truncated or absent, leading to incorrect localisation prediction. AtSAL1 illustrates this problem. The original cDNA clone isolated by (Quintero et al., 1996) encoded a 353 amino acid protein and no transit peptide. It was assumed cytosolic based on homology to ScScHAL2 and lack of transit peptide. Wilson et al. (2009) subsequently identified a 54 amino acid N-terminal extension, and Estavillo et al. (2011) experimentally confirmed plastid and mitochondrial localisation.

Database sequences derive gene prediction from algorithms applied to genome assemblies. If the AtSAL1 misannotation represents a common type of error for SAL sequences, then the propagation of this error through homology-based annotation could have created systematic misannotation of many SAL orthologues. This hypothesis is testable by examining whether "cytosolic" SAL sequences have in-frame start codons upstream of their annotated starts, with appropriate transit peptide characteristics. Such analysis could reveal the extent of annotation error in SAL across plant species.

The pH sensitivity patterns observed in this analysis provide an independent test of computational localisation predictions. If an enzyme is correctly predicted to localise to a compartment its electrostatically-derived pH optima should match. SAL clusters best ($R^2 = 0.277$, Band 3 - secondary structure at 5 Å Table 5.7) for taxa discrimination at pH 8.2 suggesting this is likely its conserved pH optima. This aligns with plastid(stroma) or peroxisome localisation. Current experimental evidence supports plastid localisation, not peroxisomal. This concordance between electrostatic pH sensitivity and predicted localisation strengthens confidence in both the localisation prediction and the biological relevance of the electrostatic patterns.

The situation is more complex for enzymes where predictions and pH sensitivity do not appear to match. Several explanations could account for these

mismatches. First, localisation prediction algorithms depend on correct identification of N-terminal targeting sequences, which requires accurate gene annotation. As discussed above, misannotation of translation start sites can truncate or eliminate targeting peptides, leading to incorrect cytosolic predictions for genuinely organellar proteins. Second, some compartments experience pH variation only under specific physiological conditions not captured by steady-state predictions. The guard cell cytosol, for example, undergoes rapid pH changes during stomatal movements and these cytosolic pH levels are atypical for the cytosols of other tissues (Bharath et al., 2024). Third, dual-targeted proteins may show pH sensitivity reflecting adaptation to whichever compartment experiences greater pH variation.

Neither localisation prediction nor pH sensitivity analysis is definitively more accurate. Rather, they provide complementary information. Localisation predictors identify targeting signals but cannot assess whether those signals are functional or whether the mature protein experiences pH variation in its destination. pH sensitivity analysis detects adaptation to dynamic pH environments but cannot distinguish between compartments with similar pH dynamics. The most reliable functional predictions emerge when both approaches converge on the same conclusion. Discordance between them should prompt experimental validation rather than automatic acceptance of either prediction. Notably subcellular pH measurements in the literature are sparse and few species have been tested, and pH sensitivity analysis demonstrated in this chapter could provide a framework for putative testing of subcellular pH conservation and generate hypothesis for pH determination in specific species given discordance may suggest greater inter-species subcellular pH variation.

This framework suggests a practical application where enzymes showing discordant predicted localisation-pH are candidates for re-examination of their gene annotations, or potentially subcellular pH determination. The combination of strong support for an optimal pH of 8.2 in SAL with such a high rate of predicted non-plastid localisation provides a rationale for systematic re-evaluation of targeting sequences in SAL orthologues.

pH sensitivity and functional adaptation

The analysis of electrostatic clustering across different pH values shows differences between paralogues that provide insight into their functional contexts and potential regulatory mechanisms. Most of the enzymes showed little variation in their ability to cluster by taxa across the pH ranges tested (Tables 5.5, 5.6, 5.7) with the exception of Algal AHL (Table 5.12). While these results

were statistically significant, the effect sizes (by R^2 values) were relatively small (Table 5.3.3). While I'll hypothesise based on the premise that pH optima (and therefore sensitivity) can be inferred across a clade, this will require experimental validation testing enzyme kinetics of predicted pH sensitive and insensitive representative homologues across a range of pH levels.

SAL showed consistent pH-independent clustering (Table 5.7), with electrostatic distances at different pH values producing statistically significant clustering.

The plastid stroma experiences large pH fluctuations during diurnal photosynthetic cycles (Aranda Sicilia et al., 2021; Trinh and Masuda, 2022; Werdan et al., 1975). An enzyme functioning in this compartment must maintain activity across this range, which would impose selective constraints on the pKa values of active site residues. The conserved pH-independent electrostatic clustering across SAL orthologues from diverse plant species may therefore reflect evolutionary optimisation for continuous function in a pH-variable environment.

CNP showed remarkable stability (Table 5.6) across pH conditions with electrostatic distances calculated at different pH values clustered together rather than separating by pH, indicating that CNP active sites maintain consistent electrostatic properties across the tested range. This pH insensitivity, combined with CNP's distinct positioning in the electrostatic similarity network relative to SAL and AHL, raises questions about CNP's functional role and substrate preference.

Chapter 4 proposed that CNP functions as a chloroplast-localised enzyme associated with photosystem II in developing tissues, potentially involved in regulating metabolic flux during growth or fine-tuning retrograde PAP signalling. The electrostatic data add complexity to this picture. CNP's substantial electrostatic distance from both SAL and AHL suggests that its active site environment differs meaningfully from the other paralogues. This electrostatic distance likely indicates a different substrate, different kinetic requirements, a distinct functional niche within the chloroplast, or a combination thereof.

If CNP does act on PAP, its pH-insensitive electrostatics might provide constitutive baseline activity that operates independently of the pH fluctuations experienced during photosynthetic cycles. However, the degree of electrostatic distance between CNP and then SAL and AHL in the similarity network suggestion caution in assuming shared substrate specificity based solely on sequence-derived classification. Experimental characterisation of CNP substrate preference and kinetic parameters would clarify whether the electrostatic differences reflect optimisation for PAP hydrolysis under different conditions or adapta-

tion to an alternative substrate entirely.

The most informative pattern came from comparing AHL sub-clades. AHL- α exhibited the strongest pH sensitivity of any group analysed, with Band 1 explaining 7.8% of variance by pH condition. AHL- β , in contrast, showed no significant pH clustering in any frequency band. This divergence between sub-clades that arose from a relatively recent gene duplication event provides direct evidence for functional specialisation following duplication.

The strong pH sensitivity of AHL- α is particularly intriguing because it suggests a potential mechanism for post-translational regulation. Unlike SAL, where redox regulation is well established, no regulatory mechanism has been characterised for AHL paralogues. If AHL- α activity is modulated by physiological pH changes, this would represent a previously unrecognised form of regulation that could coordinate AHL- α function with cellular metabolic state.

As mentioned previously, in *A. thaliana* AHL- α (encoded by *AtAHL*) shows highest expression in guard cells (Klepikova et al., 2016; Shin et al., 2019). Guard cells undergo rapid and substantial pH changes during stomatal movements, with cytosolic pH rising by approximately 0.3–0.4 during stomatal opening as proton pumps activate and ion fluxes alter the cellular environment (Bharath et al., 2024). The combination of high AHL- α expression in a pH-dynamic cell type with possible pH sensitivity in the electrostatic analysis suggests potential pH-dependent regulation.

Despite their similar baseline electrostatic properties, the divergent pH sensitivity patterns between AHL- α and AHL- β suggest regulatory differentiation invisible to sequence comparison or static structural analysis. These sub-clades arose from the γ -polyploidy event in core eudicots (Chapter 4), yet have been retained across millions of years of eudicot evolution. Gene duplication theory predicts that duplicate retention requires either sub-functionalisation, where ancestral functions are partitioned between copies, or neo-functionalisation, where one copy acquires new functions (Birchler and Yang, 2022). The electrostatic data suggest that AHL- α and AHL- β have not diverged substantially in their baseline catalytic properties, as their overlapping mid-frequency electrostatic signatures indicate comparable active site environments for substrate binding. However, AHL- α and AHL- β may perform the same reaction but under different regulatory regimes, with AHL- α subject to pH-dependent modulation while AHL- β operates constitutively.

Tissue-specific and developmental expression patterns can explain this putative difference in regulatory regimes. If AHL- β is expressed in tissues with more stable pH environments, or at developmental stages where pH regulation differs, the two paralogues could serve the same catalytic function in different

physiological contexts. This combination of spatial separation and differential pH regulation would remove selective pressure for sequence divergence while creating selective pressure to retain duplicate copies. The loss of AHL- β specifically in Brassicaceae, including *A. thaliana* (Chapter 4), may reflect lineage-specific changes in expression patterns that made one paralogue redundant in this family.

The concentration of pH effects in low-frequency bands (primarily Bands 1–2) rather than mid-frequency bands is also informative. This pattern suggests that pH-induced changes to active site electrostatics primarily affect larger-scale features across domains or the entire protein rather than fine-scale details. The mechanism likely involves titration of surface-exposed charged residues, which would alter the overall charge distribution and electrostatic field topology without substantially changing the arrangement of residues immediately surrounding the catalytic site. If pH-dependent regulation operates through AHL- α , it may function by modulating substrate access or product release through changes in the electrostatic landscape surrounding the active site, rather than by directly affecting the catalytic residues themselves.

5.4.4 What frequency bands reveal about functional encoding

A consistent finding across multiple analyses was that Bands 3 and 4, capturing spatial wavelengths of 3.7–6.7 Å outperformed other bands for paralogue discrimination. These spatial scales correspond to features such as changes in large side chain positioning, positioning of secondary structure, or of secondary structure itself.

The pitch of an α -helix, measured as the rise per complete turn, is approximately 5.4 Å (Woolfson, 2023). This falls within the optimal frequency range for paralogue discrimination. Helices contribute to active site architecture in these phosphatases, and also contribute to net charge and side-chain positioning.

Typical separations between neighbouring side chains in structured regions are also 3–6 Å (Woolfson, 2023). The arrangement of charged and polar residues around the active site creates the electrostatic environment for substrate binding and catalysis. Reorganisation of these arrangements, such as substitution of one charged residue for another or repositioning of a helix that presents several side chains, would alter electrostatic features at the mid-frequency scale.

Coordination shells around metal ions, which are essential for nucleotide phosphatase catalysis (Imam et al., 2024), have characteristic radii in this range. The magnesium ion at the active site is coordinated by oxygen atoms at distances of approximately 2–2.5 Å, but the second shell of residues providing

these coordinating groups extends to 4–6 Å (Bazayeva et al., 2024). Changes in second-shell residues could alter the electrostatic environment of the metal without directly changing coordination geometry, or alter its positioning relative to the substrate binding pocket.

The relatively poor performance of low-frequency bands (Bands 1–2) means it is difficult to assign correlation with enzyme function or classification level. Their high levels of noise likely drown out most signals. This problem likely is likely caused by the windowing method (section 5.3.3).

The concentration of discriminatory power in mid-frequency bands requires careful interpretation given the methodological constraints of this implementation. While Bands 3 and 4 showed the strongest signal, corresponding to spatial wavelengths of 3.7–6.7 Å, this observation cannot be directly interpreted as evidence that functional differences between paralogues arise primarily at secondary structure scales.

As discussed, the Hann windowing, necessary to prevent spectral leakage from the cropped active site boundaries, attenuates signal in the lowest frequency bands. Long-wavelength features extending beyond the central grid region are suppressed, potentially obscuring functionally important large-scale electrostatic gradients. The analysis was further limited to frequencies below 0.45 Å^{-1} , filtering out the highest frequency bands entirely that correspond to atomic differences. The high frequency bands that were included (bands 5 and 6) likely contain systematic errors in side chain positioning from the underlying AF2 models that we would expect to contribute noise and degrade discriminatory power. These models achieve backbone positional accuracy within 1 Å to 2 Å for high-confidence regions, but side chain positions carry greater uncertainty that exceeds the grid resolution for finer spatial features, as has been detailed in this chapter.

Consequently, the mid-frequency bands (3 and 4) likely appear most informative not because secondary structure scales carry the most biologically relevant signal, but because these frequencies fall within a measurable range bounded by windowing artefacts below and model-derived noise above.

CNP provides a notable case study regardless of these limitations. This paralogue contains a side-loop insertion that emerged after divergence from yeast lineages and is not conserved in SAL or AHL. CNP forms the most distinct cluster in both sequence and structural similarity networks (Figure 5.18A,B), and is notably more distinct in the electrostatic similarity network (Figure 5.18C). We are unable to infer from the data in this study whether the relative positioning of the paralogues in these networks arises from mid-frequency features specifically, or would be detected in other frequency bands if the high and low

frequency limitations were addressed.

These limitations could be addressed through several refinements including reducing the windowing effects on the low frequencies through different windowing algorithms, altering the window size and potentially increasing the size of the cropped region (if computational resources allow). Using AlphaFold3 or newer structural modelling tools would improve the accuracy of underlying structural models, particularly for side chain positioning. This would extend the useful frequency range allowing structural comparisons of finer scale features such as side-chains, and potentially catalytic residue positioning.

The variance explained by paralogue identity ($R^2 = 0.335$ in Band 4) represents a lower bound on the electrostatic method's discriminatory power, for reasons beyond the methodological limitations described above. This analysis validated electrostatic classification against paralogue labels derived from sequence similarity, but sequence-based classification is itself imperfect. Enzymes assigned to the same paralogue based on sequence may have diverged functionally, while enzymes in different paralogues may have converged on similar functions.

Should some of the unexplained variance likely reflects misclassification in the sequence-based labels rather than failure of electrostatic discrimination, the true relationship between electrostatics and function would be stronger than measured. Consider an enzyme incorrectly assigned to **CNP** based on sequence might be functionally more similar to **SAL**. This enzyme would increase within-paralogue electrostatic variance and decrease the apparent explanatory power of paralogue identity, even though the electrostatic method correctly captured its functional properties.

Experimental validation against kinetic parameters rather than sequence-derived labels would resolve this ambiguity. If electrostatic distance correlates more strongly with differences in k_{cat} , K_M , or pH optimum than with sequence-based paralogue assignment, this would demonstrate that the method captures functional information beyond what sequence analysis provides. Such validation represents a critical next step for establishing electrostatic comparison as a predictive tool rather than merely a correlate of evolutionary relationships.

5.4.5 Relationship to existing methods

The electrostatic comparison approach developed in this chapter complements rather than replaces existing methods for enzyme classification. Each analytical framework captures different aspects of enzyme properties and has distinct strengths and limitations which are discussed below.

Complementarity across methods

All three comparison methods (sequence-based, structure-based, and electrostatics-based) produced qualitatively similar clustering of the nucleotide phosphatase dataset (Figure 5.18). SAL, AHL, and CNP formed distinct clusters under each method, validating that these paralogue groups represent genuine biological categories rather than artefacts of any particular analytical approach.

The clustering of Algal AHL and ancient CNP sequences with SAL was most clearly apparent in the electrostatic similarity network, despite these sequences being assigned to different paralogue groups by the sequence similarity classification. The pH-dependent variation between AHL sub-clades could not have been detected by methods that do not explicitly model electrostatics (which includes environmental conditions).

Each approach has different trade-offs: **Pairwise sequence comparison** is comparatively computationally efficient, it scales relatively well, and provides a stable measure of evolutionary relatedness. However, it is limited in its ability to distinguish between sequence differences that affect function. Two enzymes with 80% sequence identity might have identical kinetic properties or might differ substantially. **Structural comparison** captures the three-dimensional arrangement of atoms that determines function, but requires high-quality structural information for each enzyme being compared. With AF2, predicted structures are now available for essentially any protein sequence, but predicted structures carry uncertainties that propagate into structural comparison metrics. While structural comparison carries these higher costs than sequence similarity this study suggests that – while limited to plant nucleotide phosphatases – it did not significantly change the clustering compared to sequence similarity (Figure 5.18A,B). **Electrostatic comparison** more directly compares proteins based on physical properties that determine enzyme-substrate interactions and catalytic activity. The electrostatic field at the active site governs substrate approach, binding orientation, transition state stabilisation, and product release. Comparing electrostatic maps therefore compares the more direct physical determinants of function. However, electrostatic calculation requires structural information and carries error propagated from it, while also introducing additional computational steps and potential error sources. Importantly it uses considerable amounts of computational power significantly constricting its utility at the moment. It did, however, provide unique insights into putative functional similarity that structural similarity did not.

For routine identification of enzyme family membership, sequence comparison remains the most practical approach. For understanding how specific

mutations affect function, structural analysis provides relevant insight. For investigating subtle functional differences between closely related enzymes, particularly those that might share similar sequences and structures but differ in their environmental responses, electrostatic analysis offers unique advantages.

Comparison to PIPSA and surface-focused methods

Some of the paralogue discrimination from this study would likely have been captured by PIPSA, even with its restriction to surface charge since surface charge distributions differ between SAL, AHL, and CNP. However, PIPSA would not have revealed the scale-specific patterns that emerged from frequency band analysis. This difference would be more significant once my method has been further refined. Furthermore, PIPSA's surface focus would exclude the buried active site region where nucleotide phosphatases bind their substrate, PAP, in an extended binding pocket, this suggests that it would not be able to discriminate the clustering relationship between SAL, Algal AHL and the ancient, algal CNP sequences.

Spherical harmonic methods

The spherical harmonic approach assumes proteins are approximately globular and that their centre of mass provides a meaningful reference point. Nucleotide phosphatases have a somewhat banana-shaped architecture when viewed side-on to the active site that would map poorly onto spherical surfaces. The active site lies in a pocket rather than at a single surface location, and projecting this three-dimensional structure onto spheres would likely obscure the functional features that drive classification. Additionally, spherical harmonic methods may sacrifice spatial relationships important for distinguishing non-surface active sites.

Hodgkin and Carbo indices

These correlation-based metrics require structural superposition and compare potentials point-by-point. For closely related homologues with high structural similarity, these methods might be suitable. However, their performance degrades significantly with structural divergence, and some members of the nucleotide phosphatase family (particularly CNP) show greater structural variation. More broadly, correlation-based metrics like the Hodgkin and Carbo indices, and methods that use them such as PIPSA and MatchTope, have been

validated primarily on structurally similar homologues where sequence identity typically exceeds 30–40%. Within and below the twilight zone (approximately 20–35% sequence identity), structural alignment itself becomes unreliable (Rost, 1999). The nucleotide phosphatase dataset includes pairwise sequence identities below 15% (see Chapter 4, Results 4.3), where alignment-dependent approaches would be expected to struggle. The Fourier approach’s reduced dependence on exact spatial correspondence may explain its apparent high performance in this dataset with very broad sequence identity.

MatchTope

MatchTope focuses specifically on peptide-binding grooves and would not be directly applicable to nucleotide phosphatase active sites, but the approach could be adapted. MatchTope uses PIPSA internally and therefore inherits its alignment requirements.

These comparisons remain speculative without the direct benchmarking of each method against the same dataset. A comprehensive benchmark comparing all approaches would be valuable future work, though the implementation effort would be substantial. Notably, some of the more computationally demanding methods such as the spherical harmonic approaches may exceed the computational resources available.

5.4.6 Limitations and opportunities for refinement

Any computational method involves approximations and introduces potential error sources. Identifying the limitations of the Fourier decomposition-Sinkhorn approach is essential for appropriate interpretation of results and for guiding future methodological development.

Structure prediction stage

The entire analysis pipeline depends on protein structures from AF2. These structures carry inherent and systematic errors that are carried through into this method.

The pLDDT metric provided by AF2 predicts the local distance difference test score, which estimates backbone positioning accuracy (Jumper et al., 2021). High pLDDT values (above 90) indicate reliable backbone positions, but the metric does not directly assess side chain positioning. Even in regions with high pLDDT, backbone conformations carry errors of 1–2 Å and this error would be

larger for side chains (Jumper et al., 2021). Since electrostatic properties depend on the precise positions of charged atoms, these side chain uncertainties translate directly into electrostatic uncertainty. This is what has likely impacted the discriminatory capacity of the higher bands (5 and 6) which would be most impacted by backbone error and, in particular, side chain positioning.

AF2 models represent static structures that cannot capture conformational dynamics and are derived from training data that relies primarily on crystal structures that do not necessarily represent the proteins' state in solution. Enzymes that sample multiple conformational states during their catalytic cycle produce averaged electrostatic signatures from the Fourier analysis that may not represent any individual functional state. If different paralogues have different conformational dynamics (or conformation in solution), this could influence the electrostatic comparison in ways not captured by the static analysis. Molecular dynamics simulations could address this limitation but are not computationally tractable at scale with current technologies.

Post-translational modifications are not modelled by AF2 including phosphorylation, acetylation, and other modifications that can substantially alter local electrostatic properties. If modification patterns differ systematically between paralogues or between taxonomic groups, the current analysis would miss this variation.

The training data for AF2 is biased toward well-characterised organisms, primarily mammals, yeast, and *Escherichia coli* (Burley et al., 2023). Plants are underrepresented, and non-model plant species are particularly sparse in the training set. SAL exhibited the lowest mean pLDDT scores among the three paralogues despite being the most experimentally characterised and having a crystal structure available. This could reflect greater sequence diversity within SAL across Viridiplantae, or it could indicate that AF2's learned representations generalise less well to the SAL clade.

Modelling all structures *de novo* without templates prioritised internal consistency over absolute accuracy. Template-based modelling for proteins with available homologue structures (such as SAL proteins related to the AtSAL1 crystal structure, PDB 8F9Y) might have improved accuracy for those proteins specifically, at the cost of introducing heterogeneity between proteins better represented by experimentally derived structures and more diverse proteins. I decided to prioritise internal consistency as this is a methodological validation study establishing whether the approach works at all. Applications to specific biological questions might benefit from template-based modelling where available.

Electrostatic calculation stage

The Poisson-Boltzmann equation provides a continuum approximation of electrostatic interactions in solution (Grochowski and Trylska, 2008). This has been validated extensively for protein electrostatics but involves simplifying assumptions.

The protein interior dielectric constant of 2 and solvent dielectric constant of 78.4 used in the calculations represent average bulk properties (Schutz and Warshel, 2001). In reality, the dielectric environment varies spatially within the protein and at the protein-solvent interface. The active site, partially enclosed and containing structured water molecules, likely has a local dielectric different from either bulk protein or bulk solvent. These local variations are not captured in this model.

Specific ion effects are not modelled and therefore doesn't capture the varied environments the enzymes are found in (at both a species level, through to tissue and subcellular compartment level). The calculation treats ions through their bulk concentration contribution to ionic strength, but specific binding of metal ions (beyond the catalytic magnesium explicitly included in the structure) or the effects of particular ions present in the *in plantae* environment are not included.

The AMBER force field provides partial atomic charges for electrostatic calculations (Cornell et al., 1996), however these parameters were derived primarily from small molecule quantum mechanical calculations and may not properly represent unusual residue environments. This includes buried charged groups or residues in unusual protonation states, which may have actual charges different from the assigned values.

Ionisation state assignment for titratable residues at each pH value relies on pKa predictions which are not precise. These predictions vary by approximately 1 pH unit even for well-validated methods (Alexov et al., 2011). Near their pKa values, small prediction errors cause large differences in calculated ionisation states and therefore electrostatic properties. This like other limitations noted above this issue is systematic and likely applied stochastically across the dataset, with differences averaging out through the law of large numbers. We cannot rule out that that a particular group of sequences may have features unique to them that introduce a systemic bias. This itself would be partially mitigated by the fact that the bias would aid in clustering these enzymes together.

Most critically – as alluded to above – default solvent and dielectric values were used in this analysis without consideration of the actual environment that

the nucleotide phosphatases reside in. This environment likely spans different subcellular compartments across plants spanning Viridiplantae occupying almost every niche available for life. We have almost no knowledge of the specific subcellular environments outside of key model organisms so default values were used.

Fourier decomposition stage

The 64^3 grid provides approximately $1.5 \text{ \AA}$ resolution for typical nucleotide phosphatase active site dimensions. Features smaller than this spacing cannot be represented accurately. While this resolution is sufficient for the mid-frequency bands that carry the strongest signal, it approaches the limit for Band 6 analysis.

Inverse distance weighting interpolation generates smooth functions based on reference points (Appendix A). In regions distant from charged residues, such as the centre of a large active site cavity, the interpolation essentially extrapolates from surrounding charges. The generated values are physically plausible but may not accurately represent electrostatic features in those regions. Increasing sampling in these non-surface regions will mitigate this but would add to the computational cost.

The power spectrum representation discards phase information so two electrostatic fields with identical frequency content but different spatial arrangements of that content produce identical power spectra. For example, a protein with all positive charges on one side and all negative charges on the other would produce the same frequency content as one with charges uniformly distributed, despite their very different electrostatic landscapes. It is highly unlikely to have two electrostatic fields with extremely similar frequency content but different spatial arrangement, but it does illustrate a key limitation of this approach. The within-band optimal transport comparison partially recovers spatial information by considering where in frequency space the power is located, but complete phase relationships between frequency components cannot be recovered.

Division into discrete frequency bands is necessary but creates artificial boundaries that require optimisation. A feature which is characterised by wavelengths that straddle two bands is split between them, potentially obscuring its contribution to either. The optimisation process that consolidated initial fine-grained bands into six final bands addressed the most problematic cases, but some boundary effects remain.

The three-dimensional Hann window applied before Fourier transformation suppresses features at the boundaries of the analysis region. This sup-

pression is most severe for low-frequency features, which require larger spatial extents to be properly characterised. An example of where this might be impacting is pH sensitivity, as discussed previously. A single oscillation of a 10–20 Å wavelength feature requires most of the grid space, yet the window attenuates these exact regions of the electrostatic map. This is a challenging problem to address, and requires balancing a trade off between low-frequency feature discrimination and spectral leakage. Future testing with different window sizes, and in particular, implementing alternative windowing methods will be required but was outside of the scope of this study (demonstrating classification based of electrostatic comparisons).

The consistently weaker discriminatory power observed in Band 1 (Table 5.3) likely reflects this suppression rather than an absence of biologically meaningful long-range electrostatic features. Long-range electrostatic steering fields, documented for enzymes such as superoxide dismutase (Getzoff et al., 1992) (and likely nucleotide phosphatases), may be underestimated by this approach. This limitation is important for interpreting the biological conclusions. For example, AHL's poor PAP binding affinity may be due to substrate selectivity not at the catalytic site, but through electrostatic steering analogous to protein channel selectivity (Mironenko et al., 2024; Chen et al., 2023).

Sinkhorn comparison stage

The optimal transport calculation within each frequency band itself introduces the potential of artefacts.

Deterministic selection of top- k points by power magnitude (Appendix A preferentially retains high-power features which tend to be associated with strongly charged surface residues because electrostatic potentials fall off with distance from the charge source. Weaker features from active site interior cavities, where potentials are lower but functionally important, may be underrepresented.

When electrostatic maps exceed 12,000 significant points, the importance sampling approach retains points with probability proportional to their electrostatic potential magnitude. Since the strongest fields typically occur at solvent-exposed charged residues, this creates a bias toward surface features at the expense of non-surface regions. This was a likely issue with Approach 2 that was specifically identified but I was unable to fully address in Approach 3 at the time. This can be addressed in future iterations through a better sparse-point sampling approach that still samples sparse regions.

The transformation to frequency space partially mitigates this bias by re-

ducing the influence of spatial position. However, the initial real-space distribution still affects which features are captured. The observation that mid-frequency bands showed stronger discriminatory power than low-frequency bands may partly reflect this surface emphasis, as non-surface features would contribute more strongly to lower frequencies.

The entropy regularisation parameter ε controls the balance between computational stability and approximation accuracy (Appendix A. The Sinkhorn algorithm computes an entropy-regularised approximation to the true EMD, with smaller ε values providing closer approximation but requiring more iterations and being more prone to numerical instability. The value used provides stable convergence across all comparisons but introduces a systematic bias compared to the unregularised optimal transport distance. This is a cost that has to be accepted due to computational intractability of EMD calculations at this scale. Never-the-less, optimisation of the ε value and the acceptable maximum iterations to coverage on an approximate optimal transport plan can be done through iterative testing and a balance was never-the-less achieved in this dataset (see Approach 3 Parameter optimisation 5.3.3).

Statistical analysis stage

Validation against sequence-derived paralogue labels is somewhat circular. Demonstrating that electrostatic similarity correlates with paralogue identity partially confirms that electrostatics correlates with sequence, since paralogues are defined by sequence. The findings where electrostatics diverges from sequence expectations, particularly the Algal AHL and ancient CNP clustering with SAL, suggest that electrostatic analysis provides independent information. However, definitive validation requires experimental characterisation of enzyme kinetics to establish whether electrostatic similarity predicts functional similarity directly. In particular, testing the kinetics of Algal AHL, representative SAL sequences, and the ancient CNP sequences would confirm or overturn the findings in this study.

Validity despite limitations

These limitations are important context for interpretation but do not undermine the central findings. The consistent and statistically well supported clustering by paralogue, taxa, and localisation across multiple frequency bands (Approach 3 Biological assessment 5.3.3 and Biological analysis beyond paralogues 5.3.3) provides evidence that the method captures genuine biological signal. The discriminatory power achieved (R^2 up to 0.335 for paralogue in Band 4)

substantially exceeds what would be expected from methodological artefacts alone. The noise titration experiments (Figure 5.20) demonstrate that clustering structure degrades gradually rather than catastrophically as noise increases, indicating robust signal recovery, and a lack of impactful artefacts.

The key question is not whether the method is perfect but whether it performs better than alternatives. Direct comparison to [PIPSA](#) or spherical harmonic methods on this dataset would require substantial additional implementation effort, but several lines of reasoning suggest the Fourier approach offers advantages for enzyme putative functional classification. [PIPSA](#)'s surface restriction would miss non-surface features that likely contribute to paralogue discrimination. Spherical harmonic methods assume approximately globular proteins and project poorly for the horseshoe-shaped nucleotide phosphatase architecture, while losing three-dimensional spatial features. The Hodgkin and Carbo indices require close structural similarity, which becomes problematic for the more divergent members of this family. The Fourier approach handles all these challenges through its transformation to frequency space and use of optimal transport for comparison.

Opportunities for methodological refinement

The methodology developed here represents an initial implementation that demonstrates proof-of-concept. There are numerous refinements that could improve performance for future applications, some of which I've flagged as I've discussed the limitations, and they and others are explored in more detail below.

Adaptive active site definition. The current approach uses a fixed rectangular prism defined by the AtSAL1 active site. This works well for closely related proteins but may exclude functionally important regions in homologues with different active site geometries. An adaptive approach that defines the active site region based on each protein's binding pocket, identified through computational methods such as fpocket ([Le Guilloux et al., 2009](#)) or through alignment of predicted binding residues, would accommodate greater structural diversity. This modification would require validation to ensure comparisons remain meaningful across different active site volumes.

Alternative windowing strategies. The Hann window provides smooth tapering that prevents spectral leakage but strongly suppresses features at analysis region boundaries. A Tukey window ([Bekka and Chikouche, 2003](#)) with a small-

ler taper window would preserve more signal at boundaries while still reducing edge artefacts. Alternatively, extending the analysis region with zero-padding before windowing would shift boundary effects further from the active site core. Systematic comparison of windowing strategies for their effects on classification performance would identify optimal approaches for different applications.

Uncertainty inheritance. The current analysis produces point estimates of electrostatic similarity without confidence intervals. Inheriting uncertainty from [AF2](#) structural predictions through electrostatic calculation and Fourier decomposition would enable more rigorous statistical inference. This would require generating multiple structural models for each sequence, either through [AF2](#)'s built-in sampling or through brief molecular dynamics relaxation, and computing electrostatic similarity distributions rather than single values. While computationally more demanding, this approach would distinguish genuine biological differences from variation attributable to structural prediction uncertainty and may be appropriate for smaller datasets.

Band optimisation and combined weighted-bands. While simple band optimisation was performed in this study (see Parameter optimisation [5.3.3](#)), detailed band optimisation would likely improve classification accuracy. This can be extended by combining the contribution of all relevant bands and weighting them against a benchmark. Different bands might contribute to clustering on the same features (such as paralogue) but do so by capturing different features at different scales. By combining these band distances into a single metric we can get a more holistic distance measure with contributions from any specified bands. Bands should not be weighted evenly as different spatial scales have different contributions to feature clustering and different noise levels. This would required careful iterative testing on a per feature basis. We could likely use the [PERMANOVA](#) analysis results to inform band weightings by scoring against a benchmark. It's not clear how generalisable this would be to other protein families, if band contribution to any given feature is not consistent with other protein families this would need to be tuned on a per-protein family basis. This would undermine the utility of this method for classification unless a sizeable subset of proteins have already been classified and can be used to tune band boundaries and weightings.

5.4.7 Generalisability beyond nucleotide phosphatases

The nucleotide phosphatases provided a convenient test system because they have clear paralogue distinctions, known phylogenetic relationships, and sufficient structural diversity to challenge the method. Whether the approach generalises to other enzyme families depends on several factors, the most notable of which are discussed below.

Factors favouring generalisation

The electrostatic approach from this study makes few assumptions specific to nucleotide phosphatases. Fourier decomposition and optimal transport are general and established mathematical frameworks applicable to any three-dimensional scalar field. The active site cropping approach could be adapted to any enzyme with a definable active site region. Alternatively for smaller proteins or smaller datasets the entire proteins' electrostatic maps could be used for comparisons. The frequency bands were chosen on theoretical grounds related to the scales of protein structural features (atomic, secondary structure, and global), which should be similar across protein families.

Factors limiting generalisation

Some aspects of the methodology were optimised specifically for nucleotide phosphatases. The active site dimensions (approximately $38 \times 36 \times 40$ Å with padding) match this enzyme family and would need adjustment for larger or smaller active sites. The grid resolution of 64^3 was chosen to balance computational resources with the accuracy of AF2 predictions for proteins of this size. Substantially larger active sites (or entire proteins) would likely require larger grids.

The frequency band scheme was empirically optimised against paralogue classification in this specific family. Other families might show different scale-dependent patterns. For example, enzymes where substrate specificity depends on a single key residue might show stronger signal in high-frequency bands, while enzymes with extended binding channels might show stronger signal in low-frequency bands.

Recommended validation approach

Before applying this method to a new enzyme family, I recommend a systematic validation similar to that performed here. Specifically, we should test whether

paralogue or functional class labels (if known) cluster significantly across multiple frequency bands using PERMANOVA. If no clustering emerges, the method may not capture the relevant functional differences for that family. If clustering is observed, the pattern of which bands show strongest discrimination will guide interpretation and may suggest modifications to band boundaries or weights. Testing an enzyme family with well characterised catalytic function functions would explore how well this method is able to capture the core enzyme function for reliable classification.

5.4.8 Implications for enzyme classification

The successful classification of nucleotide phosphatase paralogues based on active site electrostatic fields supports the broader proposition that electrostatics represents an underutilised level of functional abstraction for enzyme classification. Sequence similarity remains the dominant paradigm for inferring enzyme function, yet the results here suggest that enzymes sharing sequence similarity may diverge substantially in their electrostatic properties, while enzymes with different sequences may converge on similar electrostatic solutions.

The pH-sensitivity analysis (see Discussion pH sensitivity and functional analysis 5.4.3) particularly illustrates this point. The observation that AHL- α and AHL- β sub-clades show statistically significantly different pH responses despite their sequence similarity suggests that electrostatic classification captures functional divergence inaccessible to sequence-based methods. This has practical implications for predicting enzyme behaviour across physiological pH ranges and validating predicted localisation.

The integration of this electrostatic approach with the sequence-based methods developed in earlier chapters (Chapters 3 and 4) offers potential for improved functional annotation. Sequence similarity could be used for initial clustering, phylogenetics for paralogue classification, and with electrostatic analysis refining predictions particularly in functional classification between paralogues. This work suggests that this method is highly suited for classification within a protein family, particularly where there have been duplication events.

5.4.9 Future directions

The methodology developed here opens several avenues for future research beyond the nucleotide phosphatase application.

Experimental validation of the electrostatics-function relationship

The central hypothesis underlying this methodology is that electrostatic similarity predicts functional similarity. The analyses presented here demonstrate that electrostatic similarity correlates with evolutionary relatedness (paralogue identity, taxonomic origin), which provides indirect support. However, establishing that electrostatic comparison can predict functional properties requires direct experimental testing.

A validation study would select pairs of enzymes spanning a range of electrostatic distances, express and purify them, and characterise their kinetic parameters. If electrostatic distance correlates with differences in k_{cat} , K_{M} , or pH optimum, this would establish the method as a predictive tool. Particularly informative would be cases where sequence-similar enzymes show electrostatic divergence or where sequence-divergent enzymes show electrostatic similarity. These discordant cases would test whether electrostatics provides information beyond what sequence already reveals.

Improved sampling strategies

The deterministic top- k point selection used for [Sinkhorn](#) comparison biases toward high-magnitude surface charges. Several alternative strategies could improve representation of important but weaker non-surface electrostatic charges, as has previously been discussed.

Weighted random sampling would select points with probability proportional to power magnitude rather than deterministically by rank. This approach would ensure that high-power components dominate the comparison (as they should, being more robust) while still permitting lower-magnitude features to contribute. Repeating comparisons with different random seeds would provide uncertainty estimates.

Stratified sampling across power magnitude bins would sample the full dynamic range. For example, requiring that 20% of selected points come from each quintile of the power distribution would ensure that weak features are not excluded entirely.

Hybrid approaches combining deterministic selection of the top $k/2$ points with weighted random selection of the remainder could balance reproducibility while achieving broader coverage. The highest-power features would always contribute, providing a stable baseline, while the randomly selected component would capture additional variation.

Extension to other enzyme families

As stated earlier, applying the methodology to additional enzyme families would test how well the approach generalises. Using proteins with different folds, such as RUBISCO or TIM-barrel glycosidases, would test whether the frequency band definitions and optimal weightings transfer across protein architectures. The physical principles underlying the method should apply regardless of fold, but the specific bands carrying the strongest signal might differ.

Enzyme families with well-characterised functional variation, such as cytochrome P450s with known substrate specificities or protein kinases with defined downstream targets, would enable direct testing of whether electrostatic classification predicts functional phenotypes.

Extension to whole-protein comparison

The current implementation focuses on active site regions to reduce computational cost and increase signal-to-noise ratio. Extension to whole-protein electrostatic comparison would enable analysis of protein-protein interaction surfaces and allosteric communication networks. This would require addressing the increased computational demands and the greater structural variation expected across entire protein surfaces. As discussed previously, this could be done through reduced point sampling and grid resolution (but would reduce fine-level discriminatory power, a smaller dataset could be used, or the code could be rewritten to run in parallel on a supercomputer).

pH prediction and optimisation

The pH sensitivity analysis treated pH as a categorical grouping variable. Dedicated development could enhance this capability toward quantitative pH optimum prediction.

A training set of enzymes with experimentally determined pH optima could calibrate the relationship between frequency content changes across pH and optimal activity. If low-frequency changes correlate with pH sensitivity while mid-frequency features are pH-stable, this pattern could be used to predict pH optima for uncharacterised enzymes.

Such a tool would have practical applications in biotechnology, where enzyme pH tolerance is often a key consideration for industrial applications.

5.4.10 Conclusion

This chapter introduced a novel Fourier decomposition with Sinkhorn optimal transport for distance determination approach for comparing the electrostatic properties of enzyme active sites. By separating electrostatic fields into frequency components at different spatial scales, the method addresses noise and interpretability challenges that impacted direct comparison approaches. The use of the Sinkhorn algorithm applied a novel solution to efficiently calculate an all-versus-all distance measure that satisfies triangular equality (see section 5.3.1).

Applied to plant nucleotide phosphatases, the method successfully classified paralogues with the strongest signal in mid-frequency bands corresponding to spatial wavelengths of 3.7–6.7 Å. These scales match characteristic distances in protein secondary structure, suggesting that functional divergence between paralogues can be detected from reorganisation at this level rather than from gross charge redistribution or individual residue changes.

The analysis revealed unexpected evolutionary relationships, including the electrostatic similarity between Algal AHL and SAL and the distinct positioning of CNP from ancient eukaryotic lineages. It detected pH-dependent clustering patterns that provide insight into environmental adaptation, with divergent sensitivity between AHL sub-clades supporting neo/sub-functionalisation following gene duplication.

The methodology developed here represents the culmination of analytical approaches refined across this thesis. Chapters 3 and 4 demonstrated that static structural models, while powerful, cannot fully capture the physical determinants of protein function, and showed that SSN's coupled with molecular phylogenies can effectively classify by paralogues but cannot distinguish neutral sequence divergence from functional evolution. The electrostatic approach addresses both limitations by directly comparing the physical properties that govern enzyme-substrate interactions, while revealing functional divergence invisible to sequence analysis. This approach has putatively addressed questions about the evolution of plant nucleotide phosphatases, and the functional relationships between paralogous clades. This likely presents a strong method for family wide functional classification of plant enzyme families, subjected to frequent duplication events through plant evolution.

The unexplained variance of approximately 27% represents both a challenge and an opportunity. This residual likely encodes unmeasured biological features – likely variation in catalytic parameters. Experimental validation could establish this approach as a predictive tool for enzyme functional classifica-

tion. By extending the traditional sequence-structure-function paradigm to explicitly incorporate electrostatics, this work provides a more direct window into the physical basis of enzymatic catalysis. The challenge is finding, and applying the approach to, a suitably well characterised enzyme family. This approach works well through scale but validation by catalytic characterisation under the same conditions of hundreds of related enzymes is daunting.

The practical applications demonstrated here, from identifying annotation errors through pH sensitivity patterns to predicting functional consequences of paralogue divergence, illustrate how electrostatic analysis can generate testable hypotheses for experimental validation. The methodology provides a foundation for investigating enzyme functional evolution that complements and extends existing sequence-based and structure-based approaches.

Appendix A

Computational Pipeline

This appendix details the technical implementation of the computation pipeline.

Sparse field extraction

Sparse point sampling was implemented to ensure that regions with stronger electrostatic fields are preferentially retained in the sparse representation. The sampling probability for each point i is proportional to its electrostatic potential magnitude:

$$p_i = \frac{|\phi_i|}{\sum_{j=1}^N |\phi_j|} \quad (\text{A.1})$$

where p_i is the probability of retaining point i , ϕ_i is the electrostatic potential at point i in units of kT/e , $|\phi_i|$ denotes the absolute value (magnitude) of the potential, and the summation runs over all N significant points extracted from the electrostatic map. This weighted sampling strategy ensures that regions with stronger electrostatic fields are preferentially retained in the sparse representation.

Following sparse point extraction preprocessing the sparse, irregularly distributed points to a regular grid is necessary. Fourier transformation requires data on a regular grid, but the sparse extraction process produces points scattered throughout three-dimensional space according to the protein's charge distribution. This irregular sampling poses particular challenges in regions like binding pocket cavities, where the original electrostatic field may have sparse coverage due to the absence of nearby charged residues.

The implementation employs inverse distance weighting (IDW) interpolation to reconstruct a complete field representation on a 64^3 grid. IDW was chosen over alternatives such as radial basis function (RBF) interpolation for several reasons. IDW is computationally efficient as it does not require solving systems of equations, which becomes prohibitive for thousands of source

points. The quadratic distance falloff provides natural locality where nearby points dominate the interpolation while distant points contribute minimally. RBF interpolators can introduce oscillations in regions with sparse data coverage, which is problematic for active site cavities where the original electrostatic field may have gaps. Finally, IDW's behaviour is predictable and does not require kernel parameter selection.

For each grid point g , the interpolated electrostatic potential $\phi(g)$ combines contributions from all N source points weighted by their proximity:

$$\phi(g) = \frac{\sum_{i=1}^N w_i \phi_i}{\sum_{i=1}^N w_i} \quad (\text{A.2})$$

where ϕ_i is the electrostatic potential at source point i and w_i is the interpolation weight for that point.

The weight for each source point decreases quadratically with distance:

$$w_i = \frac{1}{d_i^2 + \varepsilon} \quad (\text{A.3})$$

where $d_i = \|g - r_i\|_2$ represents the Euclidean distance between the grid point g and the source data point position r_i , and $\varepsilon = 10^{-10}$ is a small constant that prevents division by zero when a grid point coincides exactly with a data point.

Optimal transport Mathematically, a matrix K captures the cost structure modified by the regularisation parameter ε . The algorithm maintains two scaling vectors, $\mathbf{u}$ and $\mathbf{v}$, which are iteratively updated to satisfy the marginal constraints. At each iteration:

$$\mathbf{u} = \mathbf{w}_1 \oslash (K\mathbf{v}) \quad (\text{A.4})$$

$$\mathbf{v} = \mathbf{w}_2 \oslash (K^T \mathbf{u}) \quad (\text{A.5})$$

where $\mathbf{w}_1$ and $\mathbf{w}_2$ are the source and target weight distributions respectively, K^T denotes the transpose of K , and $\oslash$ denotes element-wise division.

The implementation uses the OTT-JAX library (Cuturi et al., 2022) for GPU-accelerated optimal transport computations. This library provides numerically stable Sinkhorn solvers using log-sum-exp operations, which prevent numerical underflow that can occur with direct matrix scaling at small ε values. GPU acceleration is essential for this analysis; CPU execution would require

years rather than days for the full comparison matrix.

Distance calculation After the Sinkhorn algorithm converges, the regularised optimal transport cost quantifies the dissimilarity between the two frequency distributions:

$$\text{OT}_\varepsilon(w_1, w_2) = \langle C, T^* \rangle - \varepsilon H(T^*) \quad (\text{A.6})$$

where T^* represents the converged optimal transport plan, $\langle C, T^* \rangle$ is the transport cost under this plan, and $H(T^*)$ is its entropy.

The distance was calculated for each band, as well as an aggregate distance across all six frequency bands:

$$d_{\text{total}} = \sum_{b=1}^6 w_b \cdot d_b^{\text{OT}} \quad (\text{A.7})$$

where w_b is the weight assigned to band b and d_b^{OT} is the optimal transport distance calculated for that band.

The band weights were set based on testing with paralogue clusters:

- $w_1 = 0.10$ for band 1 (global organisation)
- $w_2 = 0.15$ for band 2 (secondary structure spacing)
- $w_3 = 0.20$ for band 3 (alpha helix pitch)
- $w_4 = 0.25$ for band 4 (beta sheet spacing)
- $w_5 = 0.20$ for band 5 (hydrogen bonds)
- $w_6 = 0.10$ for band 6 (close contacts)

These weights sum to 1, ensuring the total distance remains interpretable as a weighted average. The higher weights for bands 3, 4, and 5 likely reflect the importance of secondary structure and hydrogen bonding patterns in distinguishing paralogous enzymes, while the lower weights for the extreme bands acknowledge their reduced discriminatory power for this specific use case.

Sinkhorn algorithm implementation The Sinkhorn algorithm solves the entropy-regularised optimal transport problem through iterative matrix scaling. This approach transforms the constrained optimisation problem into a sequence of simple matrix-vector multiplications that converge to the optimal transport plan. The algorithm starts by computing the Gibbs kernel. In mathematics, a kernel is a function that takes two inputs and produces a single output measuring their relationship or similarity. The Gibbs kernel transforms

each pairwise cost into a similarity score through exponential scaling:

$$K_{ij} = \exp\left(-\frac{C_{ij}}{\varepsilon}\right) \quad (\text{A.8})$$

where C_{ij} is the transport cost between points i and j , and ε is the regularisation parameter. This transformation converts the transport costs into values between 0 and 1, where values near 1 indicate low-cost transport between similar frequency components and values near 0 indicate high-cost transport between dissimilar components.

The optimal transport plan emerges as the product of the kernel with two scaling vectors:

$$T = \text{diag}(\mathbf{u}) \cdot K \cdot \text{diag}(\mathbf{v}) \quad (\text{A.9})$$

where $\text{diag}(\mathbf{u})$ denotes a diagonal matrix with the elements of vector $\mathbf{u}$ on the diagonal, and similarly for $\mathbf{v}$.

This transformation converts the transport costs into values between 0 and 1, where values near 1 indicate low-cost transport between similar frequency components (making them easy to match) and values near 0 indicate high-cost transport between dissimilar components (making them difficult to match). The exponential function ensures smooth gradients that enable efficient optimisation. The optimal transport plan emerges as the product of the kernel with two scaling vectors:

$$T = \text{diag}(\mathbf{u}) \cdot K \cdot \text{diag}(\mathbf{v}) \quad (\text{A.10})$$

The scaling vectors are updated iteratively to satisfy the marginal constraints:

$$\mathbf{u}^{(k+1)} = \frac{\mathbf{w}_1}{K\mathbf{v}^{(k)}} \quad (\text{A.11})$$

$$\mathbf{v}^{(k+1)} = \frac{\mathbf{w}_2}{K^T\mathbf{u}^{(k+1)}} \quad (\text{A.12})$$

where $\mathbf{u}^{(k)}$ and $\mathbf{v}^{(k)}$ are the scaling vectors at iteration k , $\mathbf{w}_1$ and $\mathbf{w}_2$ are the source and target weight vectors, and the divisions are element-wise. These updates alternate between enforcing the row constraints (through $\mathbf{u}$) and column constraints (through $\mathbf{v}$).

The algorithm continues until convergence, measured by how well the marginal constraints are satisfied:

$$\|K\mathbf{v}^{(k)} \odot \mathbf{u}^{(k)} - \mathbf{w}_1\|_\infty < \tau \quad (\text{A.13})$$

where $\odot$ denotes element-wise (Hadamard) multiplication, $\|\cdot\|_\infty$ is the infinity

norm (maximum absolute value), and $\tau = 10^{-2}$ sets the convergence tolerance.

Adaptive regularisation parameter The regularisation parameter ε allows us to tune computational efficiency against solution accuracy. Too small a value leads to numerical instabilities and slow convergence, while too large a value over-smooths the transport plan, reducing discriminative power. The implementation computes ε adaptively based on the geometry of the two point clouds being compared.

A representative sample of 50 points is drawn from each cloud, yielding 2,500 pairwise distances. The median of these distances provides a robust estimate of the typical separation between frequency components:

$$d_{\text{median}} = \text{median}\{d_{ij} : i \in S_1, j \in S_2\} \quad (\text{A.14})$$

where S_1 and S_2 are the sampled point sets with $|S_1| = |S_2| = 50$, and $d_{ij} = \|x_i - x_j\|_2$ is the Euclidean distance between sampled points x_i and x_j .

The adaptive epsilon scales with the square of this median distance. The implementation applies a two-stage scaling process:

$$\varepsilon_{\text{intermediate}} = \text{clip}(0.001 \cdot d_{\text{median}}^2, 10^{-5}, 10^{-3}) \quad (\text{A.15})$$

A secondary scaling factor of 0.01 is then applied to improve discriminability between similar proteins:

$$\varepsilon_{\text{final}} = 0.01 \times \varepsilon_{\text{intermediate}} \quad (\text{A.16})$$

This yields an effective scaling factor of $\alpha = 10^{-5}$ with final values clamped to the range $[10^{-7}, 10^{-5}]$. This scaling ensures that the regularisation strength adapts to the natural scale of each comparison while maintaining sufficient sensitivity to distinguish closely related electrostatic fields.

To maintain numerical stability across all possible protein pairs, the final value is clamped to a range that was found to be reasonable through iterative testing:

$$\varepsilon_{\text{final}} = \text{clip}(\varepsilon_{\text{adaptive}}, 10^{-7}, 10^{-5}) \quad (\text{A.17})$$

This adaptive approach was essential for handling the diversity of electrostatic fields across the protein dataset, automatically adjusting to proteins with concentrated charge distributions versus those with dispersed charges.

Hann window implementation

For the 64^3 grid, the three-dimensional window function is constructed as the product of one-dimensional Hann windows along each axis:

$$w(i, j, k) = w_1(i) \cdot w_1(j) \cdot w_1(k) \quad (\text{A.18})$$

Each one-dimensional Hann window follows the standard cosine taper:

$$w_1(x) = 0.5 - 0.5 \cos\left(\frac{2\pi x}{n-1}\right) \quad (\text{A.19})$$

for $x \in \{0, 1, \dots, n-1\}$ where $n = 64$.

The windowed field is then computed as the element-wise product:

$$\phi_w(r) = \phi(r) \cdot w(r) \quad (\text{A.20})$$

Fourier transformation and physical scaling The three-dimensional discrete Fourier transform decomposes the windowed electrostatic field $\phi_w(\mathbf{r})$ into its frequency components:

$$F(\mathbf{k}) = \sum_{\mathbf{r}} \phi_w(\mathbf{r}) \exp\left(-2\pi i \frac{\mathbf{k} \cdot \mathbf{r}}{n}\right) \quad (\text{A.21})$$

where $\mathbf{r} = (r_x, r_y, r_z)$ represents spatial coordinates on the grid, $\mathbf{k} = (k_x, k_y, k_z)$ represents frequency indices in reciprocal space, n is the number of grid points per dimension (64 in this implementation), i is the imaginary unit, and $\mathbf{k} \cdot \mathbf{r}$ denotes the dot product.

The power spectrum, calculated as the squared magnitude of the Fourier coefficients:

$$P(\mathbf{k}) = |F(\mathbf{k})|^2 \quad (\text{A.22})$$

quantifies the contribution of each spatial frequency to the total electrostatic energy.

Converting the discrete frequency indices to physical units requires consideration of the grid's physical dimensions. For a grid spanning physical extent $L_x \times L_y \times L_z$ in Angstroms with n points per dimension, the frequency spacing in each dimension is:

$$\Delta f_x = \frac{1}{L_x} \text{ \AA}^{-1} \quad (\text{A.23})$$

with equivalent expressions for Δf_y and Δf_z .

After applying the [FFT](#) shift operation to centre the zero frequency, the fre-

quency coordinates become:

$$f_x[i] = \left(i - \frac{n}{2}\right) \cdot \Delta f_x \quad \text{for } i \in \{0, 1, \dots, n-1\} \quad (\text{A.24})$$

where i is the array index.

The magnitude of the three-dimensional frequency vector at each point is then:

$$|\mathbf{f}| = \sqrt{f_x^2 + f_y^2 + f_z^2} \quad (\text{A.25})$$

Sparse field extraction Each frequency band undergoes processing to extract a representative set of points for optimal transport comparison. A spherical shell mask isolates the relevant frequencies within each band. Since frequency space is three-dimensional with each point having a magnitude $|\mathbf{k}| = \sqrt{f_x^2 + f_y^2 + f_z^2}$, the mask selects all points whose magnitudes fall within the band's frequency range, creating a hollow spherical shell in frequency space. This approach groups together all frequencies corresponding to similar spatial wavelengths regardless of their orientation: A spherical shell mask $M_b(\mathbf{k})$ isolates the relevant frequencies within each band b :

$$M_b(\mathbf{k}) = \begin{cases} 1 & \text{if } f_{\min}^{(b)} \leq |\mathbf{k}| < f_{\max}^{(b)} \\ 0 & \text{otherwise} \end{cases} \quad (\text{A.26})$$

where $f_{\min}^{(b)}$ and $f_{\max}^{(b)}$ are the lower and upper frequency bounds for band b , and $|\mathbf{k}|$ is the frequency magnitude at that point.

The masked power spectrum for band b becomes:

$$P_{\text{masked}}^{(b)}(\mathbf{k}) = P(\mathbf{k}) \cdot M_b(\mathbf{k}) \quad (\text{A.27})$$

To eliminate numerical noise and focus on significant frequency components, points are retained only if they exceed a threshold based on the maximum power across all frequencies:

$$P_{\text{threshold}} = 10^{-12} \times \max_{\mathbf{k}} P(\mathbf{k}) \quad (\text{A.28})$$

This relative threshold adapts to the overall strength of each protein's electrostatic field, ensuring consistent noise filtering across proteins with different charge densities.

When more than 1,000 points in a band pass the threshold, the implementation selects the top 1,000 points ranked by power magnitude. This selection

ensures computational tractability while preserving approximately 95 % of the spectral energy within each band. The selected points form a weighted point cloud (X_b, w_b) where $X_b \in \mathbb{R}^{N \times 3}$ contains the three-dimensional frequency coordinates and the weights are normalised power values:

$$w_b^{(i)} = \frac{P_{\text{masked}}^{(b)}(\mathbf{k}_i)}{\sum_{j=1}^N P_{\text{masked}}^{(b)}(\mathbf{k}_j)} \quad (\text{A.29})$$

This weighting scheme ensures that frequency components with higher power, which contribute more to the electrostatic field's character, have proportionally greater influence in the subsequent optimal transport calculations.

COMPUTATIONAL OPTIMISATION

The computational demands of processing 2,530 proteins, requiring over 3.2 million pairwise comparisons across six frequency bands, necessitated careful optimisation of the implementation. The system leverages GPU acceleration, vectorised operations, and intelligent memory management to achieve practical run times. Never-the-less these run times would often exceed 3 days limiting the ability for iterative testing for this prototype approach.

GPU acceleration and JAX compilation

The implementation utilises JAX, a numerical computing framework that com-

bines NumPy-compatible operations with just-in-time (JIT) compilation for GPUs. GPUs were originally designed to carry out huge numbers of relatively simple calculations very quickly to display graphics for computer games. GPUs were subsequently adopted for processing vast amounts of calculations needed for machine learning and science. JAX transforms Python functions into optimised GPU kernels, achieving 50 to 100-fold speedups over CPU implementations for computationally intensive operations. Critical functions undergo JIT compilation, including the inverse distance weighting interpolation routine, FFT power spectrum computation, and Sinkhorn iterations. Compilation occurs once at first execution, with subsequent calls using cached kernels.

Efficient JIT compilation imposes the constraint that all array shapes must be statically known at compile time. This presents a challenge for optimal transport calculations where different frequency bands may yield different numbers of significant points. The implementation addresses this through fixed-size arrays with validity masks. All point clouds are padded to a maximum size of 1,000 points, with a separate validity count tracking the actual number of

meaningful points in each array. During Sinkhorn iterations, invalid points are assigned to a location outside the typical frequency range (at coordinates $(0.6, 0.6, 0.6) \text{ \AA}^{-1}$) and given zero weight, ensuring they do not influence the transport calculation. This fixed-size approach, combined with the absence of data-dependent control flow, enables significant memory optimisations and allows vectorised computation across all six frequency bands simultaneously using JAX's vmap transformation, achieving approximately 5-fold speedup over sequential band processing.

Vectorised multi-band processing

Initially I processed the six frequency bands sequentially, later I improved the code to leverage JAX to apply the Sinkhorn solver across all bands simultaneously. This vectorisation achieves approximately a n -fold speedup – where n = number of bands – compared to sequential processing by better utilising GPU parallelism and reducing kernel launch overhead.

The vectorisation requires careful memory management. Each band's point cloud is padded to a maximum size of 1,000 points (twice the typical requirement to accommodate variation), with a validity count indicating the actual number of meaningful points:

$$\text{valid}_b \in \{1, 2, \dots, 2000\} \quad (\text{A.30})$$

Invalid points beyond the validity count are masked during computation, ensuring they don't affect the transport calculations while maintaining the fixed array sizes required for efficient vectorisation. This padding strategy trades some memory overhead for substantial computational speed gains through parallelisation.

Memory hierarchy management

The implementation employs a three-tier memory hierarchy to manage the large dataset efficiently:

Tier 1 comprises disk storage for the original OpenDX files, with each file requiring between 0.5 MB and 2 MB. These files are read once during preprocessing and not accessed during the main computation phase.

Tier 2 utilises system RAM to store preprocessed band representations as NumPy arrays. Each protein's frequency representation requires:

$$\text{Memory} = 6 \text{ bands} \times 1000 \text{ points} \times 4 \text{ coordinates} \times 4 \text{ bytes} = 96 \text{ kB} \quad (\text{A.31})$$

For the complete dataset of 2,530 proteins, this totals approximately 243 MB well within the limits of the available GPU. This preprocessing step transforms the variable-sized electrostatic fields into fixed-size frequency representations, enabling efficient batch processing.

Tier 3 leverages GPU virtual random access memory (VRAM) for active computations. The batch size for concurrent processing adapts to available VRAM:

$$N_{\text{batch}} = \min \left(\frac{\text{VRAM}_{\text{available}}}{\text{Memory}_{\text{per file}}}, 150 \right) \quad (\text{A.32})$$

With 6 GB of available VRAM and approximately 500 kB required per file during processing, the system can handle batches of up to 150 protein comparisons simultaneously.

Batch processing strategy

The pairwise comparison of N proteins requires computing $N(N - 1)/2$ distances. A naive approach would transfer data to the GPU for each individual comparison, creating a severe bottleneck as the system would need to run through multiple operations for each comparison. Instead, the implementation employs a block-based batching strategy that minimises data transfers, carrying out the same number of operations per batch as a single comparison.

The proteins are partitioned into blocks of size $B = N_{\text{batch}}/2 \approx 75$. For each pair of blocks (i, j) , the system computes all pairwise distances within the block pair. When $i = j$, this yields $B(B - 1)/2$ within-block comparisons. When $i \neq j$, it produces B^2 between-block comparisons. This strategy reduces the number of GPU data transfers from $O(N^2)$ to $O(N^2/B)$. For the full dataset of 2,530 proteins with $B = 75$, this requires approximately 600 batch operations rather than 3.2 million individual transfers.

Within each batch, the GPU computes multiple protein comparisons in parallel. The Sinkhorn algorithm's matrix operations map efficiently to GPU architectures, with each iteration requiring only matrix-vector multiplications that can be highly parallelised.

Fault tolerance and checkpointing

Given the long run times for processing the complete dataset (several days even with GPU acceleration), I built multiple fault tolerance mechanisms into the code to minimise the need to rerun the dataset necessarily. This includes a checkpoint system maintains the computation state, enabling recovery from interruptions without losing progress.

The checkpoint contains the count of distance results successfully written

to disk, a boolean flag indicating whether preprocessing has completed for all proteins, and a list of any proteins that failed during preprocessing due to data quality issues or numerical problems. The checkpoint uses Python's pickle format for efficient serialisation and is saved automatically every 100 comparisons and whenever 1,000 results are written to disk.

If the computation is interrupted, the system can resume from the last checkpoint, skipping already-computed distances and continuing from where it left off. This proved essential during development when debugging required frequent restarts, and where the system was interrupted.

Quality control metrics

The system tracks preprocessing success rates to identify potential biases in the analysis. The coverage metric quantifies the fraction of possible comparisons successfully completed:

$$\text{Coverage} = \frac{N_{\text{success}}(N_{\text{success}} - 1)/2}{N_{\text{total}}(N_{\text{total}} - 1)/2} \quad (\text{A.33})$$

When coverage falls below 0.90, the system generates a warning about potential systematic bias in the results. This threshold was chosen to balance data completeness against the practical reality that some protein structures may have data quality issues preventing successful processing. In practice, the quality control approaches that were implemented earlier in the pipeline meant that no comparisons were not completed.

Failed preprocessing attempts are logged with specific error types to enable targeted debugging. Core failure modes include insufficient non-zero values in the electrostatic field (typically indicating problems with the [APBS](#) calculation), numerical instabilities during interpolation (such as with extremely sparse sampling in critical regions), and memory allocation failures (which was avoided through the careful point sampling strategies and memory handling).

Appendix B

Python structural alignment and similarity network scripts

B.1 PyMol structural alignment

```
1     import __main__
2     __main__.pymol_argv = ['pymol', '-qc']
3
4     import os
5     import pymolPy3
6
7     pm = pymolPy3.pymolPy3()
8
9     directory = '[location_of_pdb_files]'
10    output_dir = '[output_directory]'
11
12    AtSAL1_pdb = 'AtSAL1_updated.pdb' # AtSAL1 reference
13
14    """
15    Structures need to be aligned BEFORE electrostatics
16    """
17
18    # iterate over files in directory
19    for filename in os.listdir(directory):
20        basename = os.path.basename(filename)
21        f = os.path.join(directory, filename)
22
23        # checking if it is a file, create short name
24        if os.path.isfile(f):
25            shortname = basename.split('.')[0]
26
27            # load structure
28            pm(f"load_{f}")
```



```

29
30     # Load AtSAL1 reference
31     pm(f"load_{AtSAL1_pdb}")
32     pm('set_name_{AtSAL1_updated}_{AtSAL1_updated_pdb}')
33
34     pm("set_view(\
35         0.980755210,0.023148270,0.193854913,\
36         0.162131056,0.456586570,0.874780297,\
37         -0.108762056,0.889374375,0.444051415,\
38         0.000122931,0.000242302,-56.248649597,\
39         -32.444423676,-39.579414368,-9.259656906,\
40         44.643768311,67.851081848,-20.000000000)")
41
42     # set structure name
43     pdb = f"{shortname}"
44
45     # Align
46     pm(f"align_{pdb}_{AtSAL1_updated_pdb}")
47
48     # Remove AtSAL1 reference from the session
49     pm("delete_{AtSAL1_updated_pdb}")
50
51     # Save file
52     pm(f"save_{output_dir}{basename}_{pdb}")
53
54     # Clear
55     pm("reinitialize")

```

B.2 Active site coordinates

Calculation of the rectangular prism from the manually determined bottom plane are was done using the following code:

```

1  import numpy as np
2
3  corner1 = [-43.55064101, -41.4898718, -12.31982906]
4  corner2 = [-26, -45, -17]
5  corner3 = [-21.12, -35.24, -5.29]
6  corner4 = [-38.67295493, -31.73449962, -0.61338245]
7
8  distance_above = -40
9  bottom_corners = np.array([corner1, corner2, corner3, corner4])
10
11 normal_vector = np.cross(bottom_corners[1]-bottom_corners[0],
    bottom_corners[2]-bottom_corners[0])

```

```

12 normal_vector = normal_vector.astype(float) # Cast to float to
    ensure consistent data types
13 normal_vector /= np.linalg.norm(normal_vector)
14
15 top_corners = bottom_corners + distance_above * normal_vector

```

The coordinates used in PyMol to crop the active site – common to all structures – were:

```

1 PRISM_CORNERS = np.array([
2     [-43.55064101, -41.4898718, -12.31982906],
3     [-26, -45, -17],
4     [-21.12, -35.24, -5.29],
5     [-38.67295493, -31.73449962, -0.61338245],
6     [-43.8596559, -26.06520805, -25.0471327],
7     [-26.30901489, -29.57533625, -29.72730364],
8     [-21.42901489, -19.81533625, -18.01730364],
9     [-38.98196982, -16.30983587, -13.34068609]
10 ])

```

B.3 Structural similarity network construction

The code for the structural similarity network construction is detailed below:

```

1     # Define paths and parameters
2 struct_dir = "[location_of_pdb_files]"
3 output_dir = "[output_location]"
4 db_name = "foldseek_AF2_structures_all"
5
6 # Count number of PDB files for all-v-all comparison
7 pdb_files = [f for f in os.listdir(struct_dir) if f.endswith(".pdb")]
8
9 max_seq = len(pdb_files)
10 print(f"Found_{max_seq}_PDB_files_for_all-v-all_comparison")
11
12 # Create database directory structure
13 db_parent = f"{output_dir}/{db_name}_db"
14 db_loc = f"{db_parent}/{db_name}"
15
16 if not os.path.exists(db_parent):
17     os.makedirs(db_loc)
18
19 # Create foldseek database
20 os.system(f"foldseek_createdb_{struct_dir}_{db_loc}/{db_name}")
21
22 # Run search for each sequence against the database

```

```

22 for struct in pdb_files:
23     name = struct[:-4] # Remove .pdb extension
24     os.system(f"foldseek_easy-search_{struct_dir}/{struct}_{db_loc}
                _/{db_name}_{db_parent}/{name}_aln_{db_parent}/tmpFolder_--
                max-seqs_{max_seq}")
25
26 # List all alignment files (filter for _aln files only)
27 files = [os.path.join(db_parent, f) for f in os.listdir(db_parent)
           if os.path.isfile(os.path.join(db_parent, f)) and f.endswith("_aln")]
28
29 # Create CSV with all alignments
30 output_csv = f"{output_dir}/{db_name}_all-v-all.csv"
31 with open(output_csv, "w") as all_file:
32     all_file.write("query,target,fident,alnlen,msmatch,gapopen,
                    qstart,qend,tstart,tend,evalue,bits\n")
33     for f in files:
34         with open(f, "r") as ofile:
35             for line in ofile:
36                 all_file.write(line.replace("\t", ","))
37
38 print(f"All-v-all_comparison_complete. Output_saved_to:{output_csv}
      ")

```

Redundant comparisons were removed to minimise the file size. The distance matrix is comparison direction invariant, so reverse duplicates could be removed. That was done using the following code:

```

1     import pandas as pd
2
3 # Read the CSV
4 input_csv = "foldseek_AF2_structures_all_all-v-all.csv"
5 df = pd.read_csv(input_csv)
6
7 print(f"Original_records:{len(df)}")
8
9 # Remove self-comparisons
10 df = df[df['query'] != df['target']]
11 print(f"After_removing_self-comparisons:{len(df)}")
12
13 # Keep only one direction (query < target lexicographically)
14 df = df[df['query'] < df['target']]
15 print(f"After_removing_reverse_duplicates:{len(df)}")
16
17 # Save cleaned CSV
18 output_csv = "/home/mattm/projects/paper_code/20240328_colabfold/
               foldseek_AF2_structures_all_all-v-all_cleaned.csv"

```

```
19 df.to_csv(output_csv, index=False)
20
21 print(f"Cleaned CSV saved to: {output_csv}")
```

Bibliography

- Abdallah, N. A., Elsharawy, H., Abulela, H. A., Thilmony, R., Abdelhadi, A. A., and Elarabi, N. I. (2025). Multiplex CRISPR/Cas9-mediated genome editing to address drought tolerance in wheat. *GM Crops & Food*, 16(1):1–17. Publisher: Landes Bioscience _eprint: <https://doi.org/10.1080/21645698.2022.2120313>.
- Akaike, H. (1974). A New Look at the Statistical Model Identification. *IEEE Transactions on Automatic Control*, 19(6):716–723.
- Albertin, W. and Marullo, P. (2012). Polyploidy in fungi: evolution after whole-genome duplication. *Proceedings of the Royal Society B: Biological Sciences*, 279(1738):2497–2509.
- Alexov, E., Mehler, E. L., Baker, N., M. Baptista, A., Huang, Y., Milletti, F., Erik Nielsen, J., Farrell, D., Carstensen, T., Olsson, M. H. M., Shen, J. K., Warwicker, J., Williams, S., and Word, J. M. (2011). Progress in the prediction of pKa values in proteins. *Proteins: Structure, Function, and Bioinformatics*, 79(12):3260–3275. _eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1002/prot.23189>.
- Altenhoff, A. M., Glover, N. M., and Dessimoz, C. (2019). Inferring Orthology and Paralogy. In Anisimova, M., editor, *Evolutionary Genomics: Statistical and Computational Methods*, pages 149–175. Springer, New York, NY.
- Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, 215(3):403–410.
- Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, 25(17):3389–3402.
- Anandakrishnan, R., Aguilar, B., and Onufriev, A. V. (2012). H++ 3.0: automating pK prediction and the preparation of biomolecular structures

- for atomistic molecular modeling and simulations. *Nucleic Acids Research*, 40(W1):W537–W541.
- Anderson, M. J. (2001). A new method for non-parametric multivariate analysis of variance. *Austral Ecology*, 26(1):32–46. _eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1442-9993.2001.01070.pp.x>.
- Andreeva, A., Kulesha, E., Gough, J., and Murzin, A. G. (2020). The SCOP database in 2020: expanded classification of representative family and superfamily domains of known protein structures. *Nucleic Acids Research*, 48(D1):D376–D382.
- Aranda Sicilia, M. N., Sánchez Romero, M. E., Rodríguez Rosales, M. P., and Venema, K. (2021). Plastidial transporters KEA1 and KEA2 at the inner envelope membrane adjust stromal pH in the dark. *New Phytologist*, 229(4):2080–2090. _eprint: <https://nph.onlinelibrary.wiley.com/doi/pdf/10.1111/nph.17042>.
- Armbruster, U., Hertle, A., Makarenko, E., Zühlke, J., Pribil, M., Dietzmann, A., Schliebner, I., Aseeva, E., Fenino, E., Scharfenberg, M., Voigt, C., and Leister, D. (2009). Chloroplast proteins without cleavable transit peptides: Rare exceptions or a major constituent of the chloroplast proteome? *Molecular Plant*, 2(6):1325–1335.
- Armenteros, J. J. A., Salvatore, M., Emanuelsson, O., Winther, O., Heijne, G. V., Elofsson, A., and Nielsen, H. (2019). Detecting sequence signals in targeting peptides using deep learning. *Life Science Alliance*, 2(5):e201900429.
- Atkinson, H. J., Morris, J. H., Ferrin, T. E., and Babbitt, P. C. (2009). Using sequence similarity networks for visualization of relationships across diverse protein superfamilies. *PloS one*, 4(2):e4345. Publisher: Public Library of Science San Francisco, USA.
- Baker, H. G. (1965). Characteristics and modes of origin of weeds.
- Baker, N. A., Sept, D., Joseph, S., Holst, M. J., and McCammon, J. A. (2001). Electrostatics of nanosystems: Application to microtubules and the ribosome. *Proceedings of the National Academy of Sciences*, 98(18):10037–10041. Publisher: Proceedings of the National Academy of Sciences.
- Baranski, R. and Cazzonelli, C. I. (2016). Carotenoid Biosynthesis and Regulation in Plants. In *Carotenoids*, pages

- 159–189. John Wiley & Sons, Ltd. Section: 10 _eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1002/9781118622223.ch10>.
- Bazayeva, M., Andreini, C., and Rosato, A. (2024). A database overview of metal-coordination distances in metalloproteins. *Acta Crystallographica Section D: Structural Biology*, 80(5):362–376. Publisher: International Union of Crystallography.
- Bekka, R. and Chikouche, D. (2003). Effect of the window length on the EMG spectral estimation through the Blackman-Tukey method. In *Seventh International Symposium on Signal Processing and Its Applications, 2003. Proceedings.*, volume 2, pages 17–20 vol.2.
- Benjamini, Y. and Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300.
- Bevan, R. B., Bryant, D., and Lang, B. F. (2007). Accounting for gene rate heterogeneity in phylogenetic inference. *Systematic Biology*, 56(2):194–205.
- Bharath, P., Gahir, S., and Raghavendra, A. S. (2024). Cytosolic alkalinization in guard cells: an intriguing but interesting event during stomatal closure that merits further validation of its importance. *Frontiers in Plant Science*, 15. Publisher: Frontiers.
- Birchler, J. A. and Veitia, R. A. (2012). Gene balance hypothesis: Connecting issues of dosage sensitivity across biological disciplines. *Proceedings of the National Academy of Sciences*, 109(37):14746–14753. Publisher: Proceedings of the National Academy of Sciences.
- Birchler, J. A. and Yang, H. (2022). The multiple fates of gene duplications: Deletion, hypofunctionalization, subfunctionalization, neofunctionalization, dosage balance constraints, and neutral variation. *The Plant Cell*, 34(7):2466–2474.
- Blackshields, G., Sievers, F., Shi, W., Wilm, A., and Higgins, D. G. (2010). Sequence embedding for fast construction of guide trees for multiple sequence alignment. *Algorithms for Molecular Biology*, 5(1).
- Blatt, M. R. and Armstrong, F. (1993). K⁺ channels of stomatal guard cells: Absciscic-acid-evoked control of the outward rectifier mediated by cytoplasmic pH. *Planta*, 191(3):330–341.

- Blomberg, N., Gabdoulline, R. R., Nilges, M., and Wade, R. C. (1999). Classification of protein sequences by homology modeling and quantitative analysis of electrostatic similarity. *Proteins: Structure, Function, and Bioinformatics*, 37(3):379–387. _eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1002/%28SICI%291097-0134%2819991115%2937%3A3%3C379%3A%3AAID-PROT6%3E3.0.CO%3B2-K>.
- Blum, M., Andreeva, A., Florentino, L., Chuguransky, S., Grego, T., Hobbs, E., Pinto, B., Orr, A., Paysan-Lafosse, T., Ponamareva, I., Salazar, G., Bordin, N., Bork, P., Bridge, A., Colwell, L., Gough, J., Haft, D., Letunic, I., Llinares-López, F., Marchler-Bauer, A., Meng-Papaxanthos, L., Mi, H., Natale, D., Orengo, C., Pandurangan, A., Piovesan, D., Rivoire, C., Sigrist, C. A., Thanki, N., Thibaud-Nissen, F., Thomas, P., Tosatto, S. E., Wu, C., and Bateman, A. (2025). InterPro: the protein sequence classification resource in 2025. *Nucleic Acids Research*, 53(D1):D444–D456.
- Bock, R. (2010). The give-and-take of DNA: horizontal gene transfer in plants. *Trends in Plant Science*, 15(1):11–22. Publisher: Elsevier.
- Bowerman, A. F., Moore, M., Yadav, A., Zhang, J., Mortimer, M. D., Plšková, Z., Tee, E. E., Au, E. K., Collinge, D. P., Estavillo, G. M., Howitt, C. A., Chan, K. X., Rebetzke, G. J., and Pogson, B. J. (2025). Modulation of SAL retrograde signalling promotes yield and water productivity responses in dynamic field environments. *New Phytologist*, n/a(n/a). _eprint: <https://nph.onlinelibrary.wiley.com/doi/pdf/10.1111/nph.70579>.
- Branco-Price, C., Kaiser, K. A., Jang, C. J. H., Larive, C. K., and Bailey-Serres, J. (2008). Selective mRNA translation coordinates energetic and metabolic adjustments to cellular oxygen deprivation and reoxygenation in *Arabidopsis thaliana*. *The Plant Journal*, 56(5):743–755. _eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1365-313X.2008.03642.x>.
- Burley, S. K., Bhikadiya, C., Bi, C., Bittrich, S., Chao, H., Chen, L., Craig, P. A., Crichlow, G. V., Dalenberg, K., Duarte, J. M., Dutta, S., Fayazi, M., Feng, Z., Flatt, J. W., Ganesan, S., Ghosh, S., Goodsell, D. S., Green, R. K., Gur-anovic, V., Henry, J., Hudson, B. P., Khokhriakov, I., Lawson, C. L., Liang, Y., Lowe, R., Peisach, E., Persikova, I., Piehl, D. W., Rose, Y., Sali, A., Segura, J., Sekharan, M., Shao, C., Vallat, B., Voigt, M., Webb, B., Westbrook, J. D., Whetstone, S., Young, J. Y., Zalevsky, A., and Zardecki, C. (2023). RCSB Protein Data Bank (RCSB.org): delivery of experimentally-determined PDB

- structures alongside one million computed structure models of proteins from artificial intelligence/machine learning. *Nucleic Acids Research*, 51(D1):D488–D508.
- Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., and Madden, T. L. (2009). BLAST+: architecture and applications. *BMC Bioinformatics*, 10(1):421.
- Carbó, R., Leyda, L., and Arnau, M. (1980). How similar is a molecule to another? An electron density measure of similarity between two molecular structures. *International Journal of Quantum Chemistry*, 17(6):1185–1189. [_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/qua.560170612](https://onlinelibrary.wiley.com/doi/pdf/10.1002/qua.560170612).
- Cazzonelli, C. I. (2011). Carotenoids in nature: insights from plants and beyond. *Functional Plant Biology*, 38(11):833–847. Publisher: CSIRO PUBLISHING.
- Chan, K. X., Mabbitt, P. D., Phua, S. Y., Mueller, J. W., Nisar, N., Gigolashvili, T., Stroeder, E., Grassl, J., Arlt, W., Estavillo, G. M., Jackson, C. J., and Pogson, B. J. (2016). Sensing and signaling of oxidative stress in chloroplasts by inactivation of the SAL1 phosphoadenosine phosphatase. *Proceedings of the National Academy of Sciences*, 113(31):E4567–E4576. Publisher: Proceedings of the National Academy of Sciences.
- Chan, P. and Warwicker, J. (2009). Evidence for the adaptation of protein pH-dependence to subcellular pH. *BMC Biology*, 7(1):69.
- Chang, J. H., Jiao, X., Chiba, K., Oh, C., Martin, C. E., Kiledjian, M., and Tong, L. (2012). Dxo1 is a new type of eukaryotic enzyme with both decapping and 5′-3′ exoribonuclease activity. *Nature Structural & Molecular Biology*, 19(10):1011–1017. Publisher: Nature Publishing Group.
- Chen, A. Y., Brooks, B. R., and Damjanovic, A. (2023). Ion channel selectivity through ion-modulated changes of selectivity filter pKa values. *Proceedings of the National Academy of Sciences*, 120(26):e2220343120. Publisher: Proceedings of the National Academy of Sciences.
- Cheng, S.-W., Dey, T. K., and Shewchuk, J. (2016). *Delaunay Mesh Generation*. CRC Press. Google-Books-ID: oJ3SBQAAQBAJ.
- Chothia, C. and Lesk, A. M. (1986). The relation between the divergence of sequence and structure in proteins. *The EMBO Journal*, 5(4):823–826.

- Christianson, D. W. (2017). Structural and Chemical Biology of Terpenoid Cyclases. *Chemical Reviews*, 117(17):11570–11648. Publisher: American Chemical Society.
- Clark, J. W. and Donoghue, P. C. J. (2018). Whole-Genome Duplication and Plant Macroevolution. *Trends in Plant Science*, 23(10):933–945.
- Cock, P. J. A., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., and Hoon, M. J. L. D. (2009). Biopython: Freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, 25(11):1422–1423.
- Cokelaer, T., Pultz, D., Harder, L. M., Serra-Musach, J., and Saez-Rodriguez, J. (2013). BioServices: a common Python package to access biological Web Services programmatically. *Bioinformatics*, 29(24):3241–3242.
- Cooley, J. W., Lewis, P. A. W., and Welch, P. D. (1969). The Fast Fourier Transform and Its Applications. *IEEE Transactions on Education*, 12(1):27–34.
- Copp, J. N., Akiva, E., Babbitt, P. C., and Tokuriki, N. (2018). Revealing Unexplored Sequence-Function Space Using Sequence Similarity Networks. *Biochemistry*, 57(31):4651–4662.
- Cornell, W. D., Cieplak, P., Bayly, C. I., Gould, I. R., Merz, K. M., Ferguson, D. M., Spellmeyer, D. C., Fox, T., Caldwell, J. W., and Kollman, P. A. (1996). A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules J. Am. Chem. Soc. 1995, 117, 5179–5197. *Journal of the American Chemical Society*, 118(9):2309–2309. Publisher: American Chemical Society.
- Crooks, G. E., Hon, G., Chandonia, J.-M., and Brenner, S. E. (2004). WebLogo: A Sequence Logo Generator. *Genome Research*, 14(6):1188–1190. Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab.
- Cuturi, M. (2013). Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26.
- Cuturi, M., Meng-Papaxanthos, L., Tian, Y., Bunne, C., Davis, G., and Teboul, O. (2022). Optimal Transport Tools (OTT): A JAX Toolbox for all things Wasserstein. arXiv:2201.12324 [cs].

- Danchin, E. G. J., Rosso, M.-N., Vieira, P., de Almeida-Engler, J., Coutinho, P. M., Henrissat, B., and Abad, P. (2010). Multiple lateral gene transfers and duplications have promoted plant parasitism ability in nematodes. *Proceedings of the National Academy of Sciences*, 107(41):17651–17656. Publisher: Proceedings of the National Academy of Sciences.
- de Crécy-lagard, V., Amorin de Hegedus, R., Arighi, C., Babor, J., Bateman, A., Blaby, I., Blaby-Haas, C., Bridge, A. J., Burley, S. K., Cleveland, S., Colwell, L. J., Conesa, A., Dallago, C., Danchin, A., de Waard, A., Deutschbauer, A., Dias, R., Ding, Y., Fang, G., Friedberg, I., Gerlt, J., Goldford, J., Gorelik, M., Gyori, B. M., Henry, C., Hutinet, G., Jaroch, M., Karp, P. D., Kondratova, L., Lu, Z., Marchler-Bauer, A., Martin, M.-J., McWhite, C., Moghe, G. D., Monaghan, P., Morgat, A., Mungall, C. J., Natale, D. A., Nelson, W. C., O'Donoghue, S., Orengo, C., O'Toole, K. H., Radivojac, P., Reed, C., Roberts, R. J., Rodionov, D., Rodionova, I. A., Rudolf, J. D., Saleh, L., Sheynkman, G., Thibaud-Nissen, F., Thomas, P. D., Uetz, P., Vallenet, D., Carter, E. W., Weigele, P. R., Wood, V., Wood-Charlson, E. M., and Xu, J. (2022). A roadmap for the functional annotation of protein families: a community perspective. *Database*, 2022:baac062.
- del Alamo, D., Sala, D., Mchaourab, H. S., and Meiler, J. (2022). Sampling alternative conformational states of transporters and receptors with AlphaFold2. *eLife*, 11:e75751. Publisher: eLife Sciences Publications, Ltd.
- Devos, K. M., Dubcovsky, J., Dvořák, J., Chinoy, C. N., and Gale, M. D. (1995). Structural evolution of wheat chromosomes 4A, 5A, and 7B and its impact on recombination. *Theoretical and Applied Genetics*, 91(2):282–288.
- Doamekpor, S. K., Gozdek, A., Kwasnik, A., Kufel, J., and Tong, L. (2020a). A novel 5'-hydroxyl dinucleotide hydrolase activity for the DXO/Rai1 family of enzymes. *Nucleic Acids Research*, 48(1):349–358.
- Doamekpor, S. K., Grudzien-Nogalska, E., Mlynarska-Cieslak, A., Kowalska, J., Kiledjian, M., and Tong, L. (2020b). DXO/Rai1 enzymes remove 5'-end FAD and dephospho-CoA caps on RNAs. *Nucleic Acids Research*, 48(11):6136–6148.
- Dodson, G. and Wlodawer, A. (1998). Catalytic triads and their relatives. *Trends in Biochemical Sciences*, 23(9):347–352. Publisher: Elsevier.
- Duoandikoetxea, J. (2024). *Fourier Analysis*. American Mathematical Society. Google-Books-ID: mpQDEQAAQBAJ.

- Eme, L., Sharpe, S. C., Brown, M. W., and Roger, A. J. (2014). On the Age of Eukaryotes: Evaluating Evidence from Fossils and Molecular Clocks. *Cold Spring Harbor Perspectives in Biology*, 6(8):a016139. Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab.
- Emms, D. M. and Kelly, S. (2019). OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biology*, 20(1):238.
- Estavillo, G. M., Crisp, P. A., Pornsiriwong, W., Wirtz, M., Collinge, D., Carrie, C., Giraud, E., Whelan, J., David, P., Javot, H., Brearley, C., Hell, R., Marin, E., and Pogson, B. J. (2011). Evidence for a SAL1-PAP Chloroplast Retrograde Pathway That Functions in Drought and High Light Signaling in Arabidopsis. *The Plant Cell*, 23(11):3992–4012.
- Felsenstein, J. (1985). Phylogenies and the Comparative Method. *The American Naturalist*, 125(1):1–15. Publisher: The University of Chicago Press.
- Freeling, M. and Thomas, B. C. (2006). Gene-balanced duplications, like tetraploidy, provide predictable drive to increase morphological complexity. *Genome Research*, 16(7):805–814. Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab.
- Fried, S. D. and Boxer, S. G. (2017). Electric Fields and Enzyme Catalysis. *Annual Review of Biochemistry*, 86(Volume 86, 2017):387–415. Publisher: Annual Reviews.
- Fukasawa, Y., Tsuji, J., Fu, S.-C., Tomii, K., Horton, P., and Imai, K. (2015). MitoFates: Improved Prediction of Mitochondrial Targeting Sequences and Their Cleavage Sites *[S]. *Molecular & Cellular Proteomics*, 14(4):1113–1126. Publisher: Elsevier.
- Gabaldón, T. and Koonin, E. V. (2013). Functional and evolutionary implications of gene orthology. *Nature Reviews Genetics*, 14(5):360–366. Publisher: Nature Publishing Group.
- Ganguly, D. R., Yu, X., Mortimer, M. D., Moore, M., Khambalkar, P., Pogson, B. J., and Gregory, B. D. (2025). Arabidopsis DXO1 is an evolutionarily diverged homolog that impacts ribosome loading and mRNA surveillance. ISSN: 2692-8205 Pages: 2025.10.29.684918 Section: New Results.

- Garcia-Moreno, B. (2009). Adaptations of proteins to cellular and subcellular pH. *Journal of Biology*, 8(11):98.
- Garriga, E., Tommaso, P. D., Magis, C., Erb, I., Mansouri, L., Baltzis, A., Laayouni, H., Kondrashov, F., Floden, E., and Notredame, C. (2019). Large multiple sequence alignments with a root-to-leaf regressive method. *Nature Biotechnology*.
- Gašparič, M. B., Lenassi, M., Gostinčar, C., Rotter, A., Plemenitaš, A., Gunde-Cimerman, N., Gruden, K., and Žel, J. (2013). Insertion of a specific fungal 3'-phosphoadenosine-5'-phosphatase motif into a plant homologue improves halotolerance and drought tolerance of plants. *PLoS ONE*, 8(12):e81872.
- Gerlt, J. A., Bouvier, J. T., Davidson, D. B., Imker, H. J., Sadkhin, B., Slater, D. R., and Whalen, K. L. (2015). Enzyme function initiative-enzyme similarity tool (EFI-EST): A web tool for generating protein sequence similarity networks.
- Getzoff, E. D., Cabelli, D. E., Fisher, C. L., Parge, H. E., Viezzoli, M. S., Banci, L., and Hallewell, R. A. (1992). Faster superoxide dismutase mutants designed by enhancing electrostatic guidance. *Nature*, 358(6384):347–351. Publisher: Nature Publishing Group.
- Gil-Mascarell, R., López-Coronado, J. M., Bellés, J. M., Serrano, R., and Rodríguez, P. L. (1999). The Arabidopsis HAL2-like gene family includes a novel sodium-sensitive phosphatase. *Plant Journal*.
- Gligorijević, V., Renfrew, P. D., Kosciolk, T., Leman, J. K., Berenberg, D., Vatanen, T., Chandler, C., Taylor, B. C., Fisk, I. M., Vlamakis, H., Xavier, R. J., Knight, R., Cho, K., and Bonneau, R. (2021). Structure-based protein function prediction using graph convolutional networks. *Nature Communications*, 12(1):3168. Publisher: Nature Publishing Group.
- Gonatopoulos-Pournatzis, T., Dunn, S., Bounds, R., and Cowling, V. (2011). RAM/Fam103a1 Is Required for mRNA Cap Methylation. *Molecular Cell*, 44(4):585–596.
- Good, A. C. (1992). The calculation of molecular similarity: alternative formulas, data manipulation and graphical display. *Journal of Molecular Graphics*, 10(3):144–151.
- Goodstein, D. M., Shu, S., Howson, R., Neupane, R., Hayes, R. D., Fazo, J., Mitros, T., Dirks, W., Hellsten, U., Putnam, N., and Rokhsar, D. S. (2012). Phytozome: A comparative platform for green plant genomics. *Nucleic Acids Research*, 40(D1):D1178–D1186.

- Grabov, A. and Blatt, M. R. (1999). A Steep Dependence of Inward-Rectifying Potassium Channels on Cytosolic Free Calcium Concentration Increase Evoked by Hyperpolarization in Guard Cells¹. *Plant Physiology*, 119(1):277–288.
- Grochowski, P. and Trylska, J. (2008). Continuum molecular electrostatics, salt effects, and counterion binding—A review of the Poisson–Boltzmann theory and its modifications. *Biopolymers*, 89(2):93–113. _eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1002/bip.20877>.
- Gy, I., Gasciolli, V., Lauressergues, D., Morel, J. B., Gombert, J., Proux, F., Proux, C., Vaucheret, H., and Mallory, A. C. (2007). Arabidopsis FIERY1, XRN2, and XRN3 are endogenous RNA silencing suppressors. *Plant Cell*, 19(11):3451–3461.
- Hampl, V., Hug, L., Leigh, J. W., Dacks, J. B., Lang, B. F., Simpson, A. G. B., and Roger, A. J. (2009). Phylogenomic analyses support the monophyly of Excavata and resolve relationships among eukaryotic “supergroups”. *Proceedings of the National Academy of Sciences*, 106(10):3859–3864. Publisher: Proceedings of the National Academy of Sciences.
- Harris, C. R., Millman, K. J., Walt, S. J. v. d., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., Kerkwijk, M. H. v., Brett, M., Haldane, A., Río, J. F. d., Wiebe, M., Peterson, P., Gérard-Marchant, P., Sheppard, K., Reddy, T., Weckesser, W., Abbasi, H., Gohlke, C., and Oliphant, T. E. (2020). Array programming with NumPy. *Nature*, 585(7825):357–362.
- Harris, F. (1978). On the use of windows for harmonic analysis with the discrete Fourier transform. *Proceedings of the IEEE*, 66(1):51–83.
- Healy, J. and McInnes, L. (2024). Uniform manifold approximation and projection. *Nature Reviews Methods Primers*, 4(1):82. Publisher: Nature Publishing Group.
- Heldt, H. W., Werdan, K., Milovancev, M., and Geller, G. (1973). Alkalization of the chloroplast stroma caused by light-dependent proton flux into the thylakoid space. *Biochimica et Biophysica Acta (BBA) - Bioenergetics*, 314(2):224–241.
- Hie, B. L., Yang, K. K., and Kim, P. S. (2022). Evolutionary velocity with protein language models predicts evolutionary dynamics of diverse proteins. *Cell Systems*, 13(4):274–285.e6. Publisher: Elsevier.

- Hoang, D. T., Chernomor, O., Haeseler, A. V., Minh, B. Q., and Vinh, L. S. (2018). UFBoot2: Improving the ultrafast bootstrap approximation. *Molecular Biology and Evolution*, 35(2):518–522.
- Hodgkin, E. E. and Richards, W. G. (1987). Molecular similarity based on electrostatic potential and electric field. *International Journal of Quantum Chemistry*, 32(S14):105–110. _eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1002/qua.560320814>.
- Hou, X., Alagoz, Y., Welsch, R., Mortimer, M. D., Pogson, B. J., and Cazzonelli, C. I. (2024). Reducing PHYTOENE SYNTHASE activity fine-tunes the abundance of a cis-carotene-derived signal that regulates the PIF3/HY5 module and plastid biogenesis. *Journal of Experimental Botany*, 75(4):1187–1204.
- Huang, Y., Niu, B., Gao, Y., Fu, L., and Li, W. (2010). CD-HIT Suite: A web server for clustering and comparing biological sequences. *Bioinformatics*, 26(5):680–682.
- Humphrey, W., Dalke, A., and Schulten, K. (1996). VMD: Visual molecular dynamics. *Journal of Molecular Graphics*, 14(1):33–38.
- Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in Science and Engineering*, 9(3):90–95.
- Illergård, K., Ardell, D. H., and Elofsson, A. (2009). Structure is three to ten times more conserved than sequence—A study of structural response in protein cores. *Proteins: Structure, Function, and Bioinformatics*, 77(3):499–508. _eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1002/prot.22458>.
- Imam, I., Rautureau, G. J. P., Violot, S., Mulard, E. D., Magne, D., Ballut, L., Imam, I., Rautureau, G. J. P., Violot, S., Mulard, E. D., Magne, D., and Ballut, L. (2024). Structural and Functional Integration of Tissue-Nonspecific Alkaline Phosphatase Within the Alkaline Phosphatase Superfamily: Evolutionary Insights and Functional Implications. *Metabolites*, 14(12). Company: Multidisciplinary Digital Publishing Institute Distributor: Multidisciplinary Digital Publishing Institute Institution: Multidisciplinary Digital Publishing Institute Label: Multidisciplinary Digital Publishing Institute Publisher: publisher.
- Israelachvili, J. N. (2011). *Intermolecular and surface forces*. Academic press.
- Jiao, X., Chang, J. H., Kilic, T., Tong, L., and Kiledjian, M. (2013). A Mammalian Pre-mRNA 5' End Capping Quality Control Mechanism and an Unexpec-

- ted Link of Capping to Pre-mRNA Processing. *Molecular Cell*, 50(1):104–115. Publisher: Elsevier.
- Jiao, Y., Leebens-Mack, J., Ayyampalayam, S., Bowers, J. E., McKain, M. R., McNeal, J., Rolf, M., Ruzicka, D. R., Wafula, E., Wickett, N. J., Wu, X., Zhang, Y., Wang, J., Zhang, Y., Carpenter, E. J., Deyholos, M. K., Kutchan, T. M., Chanderbali, A. S., Soltis, P. S., Stevenson, D. W., McCombie, R., Pires, J. C., Wong, G. K. S., Soltis, D. E., and dePamphilis, C. W. (2012). A genome triplication associated with early diversification of the core eudicots. *Genome Biology*, 13(1):R3.
- Jones, D., Allen, J. E., Yang, Y., Drew Bennett, W. F., Gokhale, M., Moshiri, N., and Rosing, T. S. (2022). Accelerators for Classical Molecular Dynamics Simulations of Biomolecules. *Journal of Chemical Theory and Computation*, 18(7):4047–4069. Publisher: American Chemical Society.
- Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Židek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., Petersen, S., Reiman, D., Clancy, E., Zielinski, M., Steinegger, M., Pacholska, M., Berghammer, T., Bodenstein, S., Silver, D., Vinyals, O., Senior, A. W., Kavukcuoglu, K., Kohli, P., and Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589. Publisher: Nature Publishing Group.
- Kahler, U., Kamenik, A. S., Waibl, F., Kraml, J., and Liedl, K. R. (2020). Protein-Protein Binding as a Two-Step Mechanism: Preselection of Encounter Poses during the Binding of BPTI and Trypsin. *Biophysical Journal*, 119(3):652–666. Publisher: Elsevier.
- Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., von Haeseler, A., and Jermiin, L. S. (2017). ModelFinder: fast model selection for accurate phylogenetic estimates. *Nature Methods*, 14(6):587–589. Publisher: Nature Publishing Group.
- Kassambara, A. (2020). rstatix: Pipe-friendly framework for basic statistical tests. R package version 0.5.0.999.
- Katoh, K. and Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular biology and evolution*, 30(4):772–780. Publisher: Society for Molecular Biology and Evolution.

- Kerk, D., Mattice, J. F., Valdés-Tresanco, M. E., Noskov, S. Y., Ng, K. K.-S., and Moorhead, G. B. (2021). The origin and radiation of the phosphoprotein phosphatase (PPP) enzymes of Eukaryotes. *Scientific Reports*, 11(1):13681. Publisher: Nature Publishing Group.
- Kim, T. H., Böhmer, M., Hu, H., Nishimura, N., and Schroeder, J. I. (2010). Guard cell signal transduction network: Advances in understanding abscisic acid, CO₂, and Ca²⁺ signaling. *Annual Review of Plant Biology*, 61(1):561–591.
- Klepeis, J. L., Lindorff-Larsen, K., Dror, R. O., and Shaw, D. E. (2009). Long-timescale molecular dynamics simulations of protein structure and function. *Current Opinion in Structural Biology*, 19(2):120–127.
- Klepikova, A. V., Kasianov, A. S., Gerasimov, E. S., Logacheva, M. D., and Penin, A. A. (2016). A high resolution map of the *Arabidopsis thaliana* developmental transcriptome based on RNA-seq profiling. *The Plant Journal*, 88(6):1058–1070.
- Konaté, M. M., Plata, G., Park, J., Usmanova, D. R., Wang, H., and Vitkup, D. (2019). Molecular function limits divergent protein evolution on planetary timescales. *eLife*, 8:e39705. Publisher: eLife Sciences Publications, Ltd.
- Kryshtafovych, A., Schwede, T., Topf, M., Fidelis, K., and Moult, J. (2019). Critical assessment of methods of protein structure prediction (CASP)—Round XIII. *Proteins: Structure, Function, and Bioinformatics*, 87(12):1011–1020. _eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1002/prot.25823>.
- Kryshtafovych, A., Schwede, T., Topf, M., Fidelis, K., and Moult, J. (2021). Critical assessment of methods of protein structure prediction (CASP)—Round XIV. *Proteins: Structure, Function, and Bioinformatics*, 89(12):1607–1617. _eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1002/prot.26237>.
- Kwasnik, A., Wang, V. Y.-F., Krzysztan, M., Gozdek, A., Zakrzewska-Placzek, M., Stepniak, K., Poznanski, J., Tong, L., and Kufel, J. (2019). *Arabidopsis* DXO1 links RNA turnover and chloroplast function independently of its enzymatic activity. *Nucleic Acids Research*, 47(9):4751–4764.
- Küsel, A. C., Sianoudis, J., Leibfritz, D., Grimme, L. H., and Mayer, A. (1990). The dependence of the cytoplasmic pH in aerobic and anaerobic cells of the green algae *Chlorella fusca* and *Chlorella vulgaris* on the pH of the medium as determined by ³¹P in vivo NMR spectroscopy. *Archives of Microbiology*, 153(3):254–258.

- Le, S. Q. and Gascuel, O. (2010). Accounting for solvent accessibility and secondary structure in protein phylogenetics is clearly beneficial. *Systematic Biology*, 59(3):277–287.
- Le, S. Q., Lartillot, N., and Gascuel, O. (2008). Phylogenetic mixture models for proteins. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 363(1512):3965–3976.
- Le Guilloux, V., Schmidtke, P., and Tuffery, P. (2009). Fpocket: An open source platform for ligand pocket detection. *BMC Bioinformatics*, 10(1):168.
- Leebens-Mack, J. H., Barker, M. S., Carpenter, E. J., Deyholos, M. K., Gitzendanner, M. A., Graham, S. W., Grosse, I., Li, Z., Melkonian, M., Mirarab, S., Porsch, M., Quint, M., Rensing, S. A., Soltis, D. E., Soltis, P. S., Stevenson, D. W., Ullrich, K. K., Wickett, N. J., DeGironimo, L., Edger, P. P., Jordon-Thaden, I. E., Joya, S., Liu, T., Melkonian, B., Miles, N. W., Pokorný, L., Quigley, C., Thomas, P., Villarreal, J. C., Augustin, M. M., Barrett, M. D., Baucom, R. S., Beerling, D. J., Benstein, R. M., Biffin, E., Brockington, S. F., Burge, D. O., Burris, J. N., Burris, K. P., Burtet-Sarramegna, V., Caicedo, A. L., Cannon, S. B., Çebi, Z., Chang, Y., Chater, C., Cheeseman, J. M., Chen, T., Clarke, N. D., Clayton, H., Covshoff, S., Crandall-Stotler, B. J., Cross, H., dePamphilis, C. W., Der, J. P., Determann, R., Dickson, R. C., Stilio, V. S. D., Ellis, S., Fast, E., Feja, N., Field, K. J., Filatov, D. A., Finnegan, P. M., Floyd, S. K., Fogliani, B., García, N., Gâteblé, G., Godden, G. T., Goh, F. Q. Y., Greiner, S., Harkess, A., Heaney, J. M., Helliwell, K. E., Heyduk, K., Hibberd, J. M., Hodel, R. G. J., Hollingsworth, P. M., Johnson, M. T. J., Jost, R., Joyce, B., Kapralov, M. V., Kazamia, E., Kellogg, E. A., Koch, M. A., Konrat, M. V., Könyves, K., Kutchan, T. M., Lam, V., Larsson, A., Leitch, A. R., Lentz, R., Li, F. W., Lowe, A. J., Ludwig, M., Manos, P. S., Mavrodiev, E., McCormick, M. K., McKain, M., McLellan, T., McNeal, J. R., Miller, R. E., Nelson, M. N., Peng, Y., Ralph, P., Real, D., Riggins, C. W., Ruhsam, M., Sage, R. F., Sakai, A. K., Scascitella, M., Schilling, E. E., Schlösser, E. M., Sederoff, H., Servick, S., Sessa, E. B., Shaw, A. J., Shaw, S. W., Sigel, E. M., Skema, C., Smith, A. G., Smithson, A., Stewart, C. N., Stinchcombe, J. R., Szövényi, P., Tate, J. A., Tiebel, H., Trapnell, D., Villegente, M., Wang, C. N., Weller, S. G., Wenzel, M., Weststrand, S., Westwood, J. H., Whigham, D. F., Wu, S., Wulff, A. S., Yang, Y., Zhu, D., Zhuang, C., Zuidhof, J., Chase, M. W., Pires, J. C., Rothfels, C. J., Yu, J., Chen, C., Chen, L., Cheng, S., Li, J., Li, R., Li, X., Lu, H., Ou, Y., Sun, X., Tan, X., Tang, J., Tian, Z., Wang, F., Wang, J., Wei, X., Xu, X., Yan, Z., Yang, F., Zhong, X., Zhou, F., Zhu, Y., Zhang, Y., Ayyampalayam, S., Barkman, T. J., Nguyen, N. p., Matasci, N., Nelson, D. R., Sayyari, E., Wafula, E. K., Walls,

- R. L., Warnow, T., An, H., Arrigo, N., Baniaga, A. E., Galuska, S., Jorgensen, S. A., Kidder, T. I., Kong, H., Lu-Irving, P., Marx, H. E., Qi, X., Reardon, C. R., Sutherland, B. L., Tiley, G. P., Welles, S. R., Yu, R., Zhan, S., Gramzow, L., Theißen, G., and Wong, G. K. S. (2019). One thousand plant transcriptomes and the phylogenomics of green plants. *Nature*, 574(7780):679–685.
- Li, W. and Godzik, A. (2006). Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, 22(13):1658–1659.
- Lin, B., Luo, X., Liu, Y., and Jin, X. (2024). A comprehensive review and comparison of existing computational methods for protein function prediction. *Briefings in Bioinformatics*, 25(4):bbae289.
- Lin, P.-Y., Huang, S.-C., Chen, K.-L., Huang, Y.-C., Liao, C.-Y., Lin, G.-J., Lee, H., and Chen, P.-Y. (2025). Analysing protein complexes in plant science: insights and limitation with AlphaFold 3. *Botanical Studies*, 66(1):14.
- Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., Verkuil, R., Kabeli, O., Shmueli, Y., Dos Santos Costa, A., Fazel-Zarandi, M., Sercu, T., Candido, S., and Rives, A. (2023). Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130.
- Liu, C., Liu, B., Zhang, Y., Jiang, F., Ren, Y., Li, S., Wang, H., and Fan, W. (2018). Ancient horizontally transferred genes in the genome of California two-spot octopus, *Octopus bimaculoides*. *Gene*, 667:34–44.
- Liu, F., Wang, S.-H., Cheewangkoon, R., and Zhao, R.-L. (2024). Uneven distribution of prokaryote-derived horizontal gene transfer in fungi: a lifestyle-dependent phenomenon. *mBio*, 16(1):e02855–24. Publisher: American Society for Microbiology.
- Liu, T., Liu, Q., Yu, Z., Wang, C., Mai, H., Liu, G., Li, R., Pang, G., Chen, D., Liu, H., Yang, J., and Tao, L.-Z. (2022). eIF4E1 Regulates Arabidopsis Embryo Development and Root Growth by Interacting With RopGEF7. *Frontiers in Plant Science*, 13. Publisher: Frontiers.
- Lynch, M. and Marinov, G. K. (2015). The bioenergetic costs of a gene. *Proceedings of the National Academy of Sciences*, 112(51):15690–15695. Publisher: Proceedings of the National Academy of Sciences.
- Mandel, J. R., Dikow, R. B., Siniscalchi, C. M., Thapa, R., Watson, L. E., and Funk, V. A. (2019). A fully resolved backbone phylogeny reveals numer-

- ous dispersals and explosive diversifications throughout the history of Asteraceae. *Proceedings of the National Academy of Sciences*, 116(28):14083–14088. Publisher: Proceedings of the National Academy of Sciences.
- Manmathan, H., Shaner, D., Snelling, J., Tisserat, N., and Lapitan, N. (2013). Virus-induced gene silencing of *Arabidopsis thaliana* gene homologues in wheat identifies genes conferring improved drought tolerance. *Journal of Experimental Botany*, 64(5):1381–1392.
- Maquat, L. E., Tarn, W.-Y., and Isken, O. (2010). The Pioneer Round of Translation: Features and Functions. *Cell*, 142(3):368–374. Publisher: Elsevier.
- Marcussen, T., Sandve, S. R., Heier, L., Spannagl, M., Pfeifer, M., THE INTERNATIONAL WHEAT GENOME SEQUENCING CONSORTIUM,, Jakobsen, K. S., Wulff, B. B. H., Steuernagel, B., Mayer, K. F. X., and Olsen, O.-A. (2014). Ancient hybridizations among the ancestral genomes of bread wheat. *Science*, 345(6194):1250092. Publisher: American Association for the Advancement of Science.
- Mariani, V., Biasini, M., Barbato, A., and Schwede, T. (2013). IDDT: A local superposition-free score for comparing protein structures and models using distance difference tests. *Bioinformatics*, 29(21):2722–2728.
- Mariani, V., Kiefer, F., Schmidt, T., Haas, J., and Schwede, T. (2011). Assessment of template based protein structure predictions in CASP9. *Proteins: Structure, Function, and Bioinformatics*, 79(S10):37–58. _eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1002/prot.23177>.
- Martínez-Silva, A. V., Aguirre-Martínez, C., Flores-Tinoco, C. E., Alejandri-Ramírez, N. D., and Dinkova, T. D. (2012). Translation Initiation Factor AteIF(iso)4E Is Involved in Selective mRNA Translation in *Arabidopsis Thaliana* Seedlings. *PLOS ONE*, 7(2):e31606. Publisher: Public Library of Science.
- Masterson, J. (1994). Stomatal Size in Fossil Plants: Evidence for Polyploidy in Majority of Angiosperms. *Science*, 264(5157):421–424.
- Mayberry, L. K., Allen, M. L., Dennis, M. D., and Browning, K. S. (2009). Evidence for Variation in the Optimal Translation Initiation Complex: Plant eIF4B, eIF4F, and eIF(iso)4F Differentially Promote Translation of mRNAs. *Plant Physiology*, 150(4):1844–1854.
- McCoy, A. J., Chandana Epa, V., and Colman, P. M. (1997). Electrostatic complementarity at protein/protein interfaces1. *Journal of Molecular Biology*, 268(2):570–584.

- McKinney, W. (2010). Data Structures for Statistical Computing in Python. In Walt, S. v. d. and Millman, J., editors, *Proceedings of the 9th Python in Science Conference*, pages 56–61.
- Mendes, M. F. d. A., de Souza Bragatte, M., Vianna, P., de Freitas, M. V., Pöhner, I., Richter, S., Wade, R. C., Salzano, F. M., and Vieira, G. F. (2022). MatchTope: A tool to predict the cross reactivity of peptides complexed with Major Histocompatibility Complex I. *Frontiers in Immunology*, 13. Publisher: Frontiers.
- Mi, H., Huang, X., Muruganujan, A., Tang, H., Mills, C., Kang, D., and Thomas, P. D. (2017). PANTHER version 11: expanded annotation data from Gene Ontology and Reactome pathways, and data analysis tool enhancements. *Nucleic Acids Research*, 45(D1):D183–D189.
- Millar, A. H., Carrie, C., Pogson, B., and Whelan, J. (2009). Exploring the Function-Location Nexus: Using Multiple Lines of Evidence in Defining the Subcellular Location of Plant Proteins. *The Plant Cell*, 21(6):1625 LP – 1631.
- Minh, B. Q., Nguyen, M. A. T., and Haeseler, A. V. (2013). Ultrafast approximation for phylogenetic bootstrap. *Molecular Biology and Evolution*, 30(5):1188–1195.
- Minh, B. Q., Schmidt, H. A., Chernomor, O., Schrempf, D., Woodhams, M. D., von Haeseler, A., and Lanfear, R. (2020). IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Molecular Biology and Evolution*, 37(5):1530–1534.
- Mirdita, M., Schütze, K., Moriwaki, Y., Heo, L., Ovchinnikov, S., and Steinegger, M. (2022). ColabFold: making protein folding accessible to all. *Nature Methods*, 19(6):679–682. Publisher: Nature Publishing Group.
- Mironenko, A., de Groot, B. L., and Kopec, W. (2024). Selectivity filter mutations shift ion permeation mechanism in potassium channels. *PNAS Nexus*, 3(7):pgae272.
- Mistry, J., Chuguransky, S., Williams, L., Qureshi, M., Salazar, G. A., Sonnhammer, E. L. L., Tosatto, S. C. E., Paladin, L., Raj, S., Richardson, L. J., Finn, R. D., and Bateman, A. (2021). Pfam: The protein families database in 2021. *Nucleic Acids Research*, 49(D1):D412–D419.
- Mohr, T., Horstman, J., Gu, Y. Q., Elarabi, N. I., Abdallah, N. A., and Thilmony, R. (2022). CRISPR-Cas9 Gene Editing of the Sal1 Gene Family in Wheat. *Plants*, 11(17):2259. Publisher: Multidisciplinary Digital Publishing Institute.

- Morris, R. J., Najmanovich, R. J., Kahraman, A., and Thornton, J. M. (2005). Real spherical harmonic expansion coefficients as 3D shape descriptors for protein binding pocket and ligand comparisons. *Bioinformatics*, 21(10):2347–2355.
- Morrison, D. A. (2006). Multiple sequence alignment for phylogenetic purposes. *Australian Systematic Botany*, 19(6):479.
- Murguía, J. R., Bellés, J. M., and Serrano, R. (1995). A salt-sensitive 3'(2'),5'-bisphosphate nucleotidase involved in sulfate activation. *Science*, 267(5195):232–234.
- Nabhan, A. R. and Sarkar, I. N. (2012). The impact of taxon sampling on phylogenetic inference: a review of two decades of controversy. *Briefings in Bioinformatics*, 13(1):122–134.
- Nagano, N., Orengo, C. A., and Thornton, J. M. (2002). One Fold with Many Functions: The Evolutionary Relationships between TIM Barrel Families Based on their Sequences, Structures and Functions. *Journal of Molecular Biology*, 321(5):741–765.
- Nakagawa, S. and Cuthill, I. C. (2007). Effect size, confidence interval and statistical significance: a practical guide for biologists. *Biological Reviews*, 82(4):591–605. [_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1469-185X.2007.00027.x](https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1469-185X.2007.00027.x).
- Nascimento, F. F., Reis, M. d., and Yang, Z. (2017). A biologist's guide to Bayesian phylogenetic analysis. *Nature Ecology & Evolution*, 1(10):1446–1454. Publisher: Nature Publishing Group.
- Notredame, C. (2010). Computing multiple sequence/structure alignments with the T-coffee package. *Current Protocols in Bioinformatics*, 4(SUPPL. 29):381–3825.
- Notredame, C., Higgins, D. G., and Heringa, J. (2000). T-coffee: a novel method for fast and accurate multiple sequence alignment1. *Journal of Molecular Biology*, 302(1):205–217.
- Obayashi, T., Aoki, Y., Tadaka, S., Kagaya, Y., and Kinoshita, K. (2018). ATTED-II in 2018: A Plant Coexpression Database Based on Investigation of the Statistical Property of the Mutual Rank Index. *Plant and Cell Physiology*, 59(1):e3–e3.

- Oberg, N., Zallot, R., and Gerlt, J. A. (2023). EFI-EST, EFI-GNT, and EFI-CGFP: Enzyme Function Initiative (EFI) Web Resource for Genomic Enzymology Tools. *Journal of Molecular Biology*, 435(14):168018.
- Olsson, M. H. M., Søndergaard, C. R., Rostkowski, M., and Jensen, J. H. (2011). PROPKA3: Consistent Treatment of Internal and Surface Residues in Empirical pKa Predictions. *Journal of Chemical Theory and Computation*, 7(2):525–537. Publisher: American Chemical Society.
- Omelchenko, M. V., Galperin, M. Y., Wolf, Y. I., and Koonin, E. V. (2010). Non-homologous isofunctional enzymes: A systematic analysis of alternative solutions in enzyme evolution. *Biology Direct*, 5(1):31.
- Onufriev, A. V. and Case, D. A. (2019). Generalized Born Implicit Solvent Models for Biomolecules. *Annual Review of Biophysics*, 48(Volume 48, 2019):275–296. Publisher: Annual Reviews.
- Paganini, J., Campan-Fournier, A., Rocha, M. D., Gouret, P., Pontarotti, P., Wajenberg, E., Abad, P., and Danchin, E. G. J. (2012). Contribution of Lateral Gene Transfers to the Genome Composition and Parasitic Ability of Root-Knot Nematodes. *PLOS ONE*, 7(11):e50875. Publisher: Public Library of Science.
- Pan, S., Li, K., Huang, W., Zhong, H., Wu, H., Wang, Y., Zhang, H., Cai, Z., Guo, H., Chen, X., and Xia, Y. (2020). *Arabidopsis* DXO1 possesses deNADding and exonuclease activities and its mutation affects defense-related and photosynthetic gene expression. *Journal of Integrative Plant Biology*, 62(7):967–983.
- Panchy, N., Lehti-Shiu, M., and Shiu, S. H. (2016). Evolution of gene duplication in plants. *Plant Physiology*, 171(4):2294–2316.
- Paradis, E. and Schliep, K. (2019). Ape 5.0: An environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics*, 35(3):526–528.
- Parfrey, L. W., Lahr, D. J. G., Knoll, A. H., and Katz, L. A. (2011). Estimating the timing of early eukaryotic diversification with multigene molecular clocks. *Proceedings of the National Academy of Sciences*, 108(33):13624–13629. Publisher: Proceedings of the National Academy of Sciences.
- Patrick, R. M. and Browning, K. S. (2012). The eIF4F and eIFiso4F Complexes of Plants: An Evolutionary Perspective. *International Journal of Genomics*, 2012(1):287814. eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1155/2012/287814>.

- Potter, S. C., Luciani, A., Eddy, S. R., Park, Y., Lopez, R., and Finn, R. D. (2018). HMMER web server: 2018 update. *Nucleic Acids Research*, 46(W1):W200–W204.
- Poudel, S., Pike, D. H., Raanan, H., Mancini, J. A., Nanda, V., Rickaby, R. E. M., and Falkowski, P. G. (2020). Biophysical analysis of the structural evolution of substrate specificity in RuBisCO. *Proceedings of the National Academy of Sciences*, 117(48):30451–30457. Publisher: Proceedings of the National Academy of Sciences.
- Qiao, X., Li, Q., Yin, H., Qi, K., Li, L., Wang, R., Zhang, S., and Paterson, A. H. (2019). Gene duplication and evolution in recurring polyploidization-diploidization cycles in plants. *Genome Biology*, 20(1):38.
- Quintero, F. J., Garciadeblás, B., and Rodríguez-Navarro, A. (1996). The SAL1 gene of arabidopsis, encoding an enzyme with 3'(2'),5'-bisphosphate nucleotidase and inositol polyphosphate 1-phosphatase activities, increases salt tolerance in yeast. *Plant Cell*, 8(3):529–537.
- R Core Team (2019). A Language and Environment for Statistical Computing.
- Radivojac, P., Clark, W. T., Oron, T. R., Schnoes, A. M., Wittkop, T., Sokolov, A., Graim, K., Funk, C., Verspoor, K., Ben-Hur, A., Pandey, G., Yunes, J. M., Talwalkar, A. S., Repo, S., Souza, M. L., Piovesan, D., Casadio, R., Wang, Z., Cheng, J., Fang, H., Gough, J., Koskinen, P., Törönen, P., Nokso-Koivisto, J., Holm, L., Cozzetto, D., Buchan, D. W. A., Bryson, K., Jones, D. T., Limaye, B., Inamdar, H., Datta, A., Manjari, S. K., Joshi, R., Chitale, M., Kihara, D., Lisewski, A. M., Erdin, S., Venner, E., Lichtarge, O., Rentzsch, R., Yang, H., Romero, A. E., Bhat, P., Paccanaro, A., Hamp, T., Kaßner, R., Seemayer, S., Vicedo, E., Schaefer, C., Achten, D., Auer, F., Boehm, A., Braun, T., Hecht, M., Heron, M., Hönigsmid, P., Hopf, T. A., Kaufmann, S., Kiening, M., Krompass, D., Landerer, C., Mahlich, Y., Roos, M., Björne, J., Salakoski, T., Wong, A., Shatkay, H., Gatzmann, F., Sommer, I., Wass, M. N., Sternberg, M. J. E., Škunca, N., Supek, F., Bošnjak, M., Panov, P., Džeroski, S., Šmuc, T., Kourmpetis, Y. A. I., van Dijk, A. D. J., Braak, C. J. F. t., Zhou, Y., Gong, Q., Dong, X., Tian, W., Falda, M., Fontana, P., Lavezzo, E., Di Camillo, B., Toppo, S., Lan, L., Djuric, N., Guo, Y., Vucetic, S., Bairoch, A., Linial, M., Babbitt, P. C., Brenner, S. E., Orengo, C., Rost, B., Mooney, S. D., and Friedberg, I. (2013). A large-scale evaluation of computational protein function prediction. *Nature Methods*, 10(3):221–227. Publisher: Nature Publishing Group.

- Rasheed, F., Markgren, J., Hedenqvist, M., and Johansson, E. (2020). Modeling to Understand Plant Protein Structure-Function Relationships—Implications for Seed Storage Proteins. *Molecules*, 25(4). Company: Multidisciplinary Digital Publishing Institute Distributor: Multidisciplinary Digital Publishing Institute Institution: Multidisciplinary Digital Publishing Institute Label: Multidisciplinary Digital Publishing Institute Publisher: publisher.
- Reiser, L., Bakker, E., Subramaniam, S., Chen, X., Sawant, S., Khosa, K., Prithvi, T., and Berardini, T. Z. (2024). The Arabidopsis Information Resource in 2024. *Genetics*, 227(1):iyae027.
- Ribeiro, A. J. M., Riziotis, I. G., Borkakoti, N., and Thornton, J. M. (2023). Enzyme function and evolution through the lens of bioinformatics. *The Biochemical Journal*, 480(22):1845–1863.
- Richard, J. P. (2019). Protein Flexibility and Stiffness Enable Efficient Enzymatic Catalysis. *Journal of the American Chemical Society*, 141(8):3320–3331. Publisher: American Chemical Society.
- Richards, T. A., Soanes, D. M., Foster, P. G., Leonard, G., Thornton, C. R., and Talbot, N. J. (2009). Phylogenomic Analysis Demonstrates a Pattern of Rare and Ancient Horizontal Gene Transfer between Plants and Fungi. *The Plant Cell*, 21(7):1897–1911.
- Richter, S., Wenzel, A., Stein, M., Gabdoulline, R. R., and Wade, R. C. (2008). webPIPSA: a web server for the comparison of protein interaction properties. *Nucleic Acids Research*, 36(suppl_2):W276–W280.
- Ritchie, D. W. and Kemp, G. J. (2000). Protein docking using spherical polar Fourier correlations. *Proteins: Structure, Function, and Bioinformatics*, 39(2):178–194. _eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1002/%28SICI%291097-0134%2820000501%2939%3A2%3C178%3A%3AAID-PROT8%3E3.0.CO%3B2-6>.
- Romero-Romero, S., Kordes, S., Michel, F., and Höcker, B. (2021). Evolution, folding, and design of TIM barrels and related proteins. *Current Opinion in Structural Biology*, 68:94–104.
- Rost, B. (1999). Twilight zone of protein sequence alignments. *Protein Engineering*, 12(2):85–94.

- Rubner, Y., Tomasi, C., and Guibas, L. J. (2000). The Earth Mover's Distance as a Metric for Image Retrieval. *International Journal of Computer Vision*, 40(2):99–121.
- Ruiz-Pernía, J. J., Świderek, K., Bertran, J., Moliner, V., and Tuñón, I. (2024). Electrostatics as a Guiding Principle in Understanding and Designing Enzymes. *Journal of Chemical Theory and Computation*, 20(5):1783–1795. Publisher: American Chemical Society.
- Ruiz-Sola, M. U., Coman, D., Beck, G., Barja, M. V., Colinas, M., Graf, A., Welsch, R., Rütimann, P., Bühlmann, P., Bigler, L., Gruissem, W., Rodríguez-Concepción, M., and Vranová, E. (2016). Arabidopsis GERANYLGERANYL DIPHOSPHATE SYNTHASE 11 is a hub isozyme required for the production of most photosynthesis-related isoprenoids. *New Phytologist*, 209(1):252–264. _eprint: <https://nph.onlinelibrary.wiley.com/doi/pdf/10.1111/nph.13580>.
- Sajeev, N., Bai, B., and Bentsink, L. (2019). Seeds: A Unique System to Study Translational Regulation. *Trends in Plant Science*, 24(6):487–495. Publisher: Elsevier.
- Salazar-Díaz, K., Aquino-Luna, M., Hernández-Lucero, E., Nieto-Rivera, B., Pulido-Torres, M. A., Jorge-Pérez, J. H., Gavilanes-Ruiz, M., and Dinkova, T. D. (2021). Arabidopsis thaliana eIF4E1 and eIF(iso)4E Participate in Cold Response and Promote Translation of Some Stress-Related mRNAs. *Frontiers in Plant Science*, 12. Publisher: Frontiers.
- Schnoes, A. M., Brown, S. D., Dodevski, I., and Babbitt, P. C. (2009). Annotation Error in Public Databases: Misannotation of Molecular Function in Enzyme Superfamilies. *PLOS Computational Biology*, 5(12):e1000605. Publisher: Public Library of Science.
- Schutz, C. N. and Warshel, A. (2001). What are the dielectric “constants” of proteins and how to validate electrostatic models? *Proteins: Structure, Function, and Bioinformatics*, 44(4):400–417. _eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1002/prot.1106>.
- Schwarz, G. (1978). Estimating the Dimension of a Model. *The Annals of Statistics*, 6(2):461–464.
- Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003). Cytoscape: A software Environment for integrated models of biomolecular interaction networks. *Genome Research*, 13(11):2498–2504.

- Sharwood, R. E. (2017). Engineering chloroplasts to improve Rubisco catalysis: prospects for translating improvements into food and fiber crops. *New Phytologist*, 213(2):494–510. _eprint: <https://nph.onlinelibrary.wiley.com/doi/pdf/10.1111/nph.14351>.
- Shavrukov, Y., Kurishbayev, A., Jatayev, S., Shvidchenko, V., Zotova, L., Koeke-moer, F., de Groot, S., Soole, K., and Langridge, P. (2017). Early Flowering as a Drought Escape Mechanism in Plants: How Can It Aid Wheat Production? *Frontiers in Plant Science*, 8. Publisher: Frontiers.
- Shen, J., Zeng, Y., Zhuang, X., Sun, L., Yao, X., Pimpl, P., and Jiang, L. (2013). Organelle pH in the Arabidopsis Endomembrane System. *Molecular Plant*, 6(5):1419–1437. Publisher: Elsevier.
- Shimodaira, H. (2002). An approximately unbiased test of phylogenetic tree selection. *Systematic Biology*, 51(3):492–508.
- Shin, D. J., Min, J. H., Nguyen, T. V., Kim, Y. M., and Kim, C. S. (2019). Loss of Arabidopsis Halotolerance 2-like (AHL), a 3'-phosphoadenosine-5'-phosphate phosphatase, suppresses insensitive response of Arabidopsis thaliana ring zinc finger 1 (atrzf1) mutant to abiotic stress. *Plant Molecular Biology*, 99(4-5):363–377.
- Singh, R. K., Svystun, T., AlDahmash, B., Jönsson, A. M., and Bhalerao, R. P. (2017). Photoperiod- and temperature-mediated control of phenology in trees – a molecular perspective. *New Phytologist*, 213(2):511–524. _eprint: <https://nph.onlinelibrary.wiley.com/doi/pdf/10.1111/nph.14346>.
- Sonenberg, N. and Hinnebusch, A. G. (2009). Regulation of Translation Initiation in Eukaryotes: Mechanisms and Biological Targets. *Cell*, 136(4):731–745. Publisher: Elsevier.
- Song, N., Joseph, J. M., Davis, G. B., and Durand, D. (2008). Sequence similarity network reveals common ancestry of multidomain proteins. *PLoS Computational Biology*, 4(5):e1000063–e1000063.
- Spreitzer, R. J. and Salvucci, M. E. (2002). RUBISCO: Structure, Regulatory Interactions, and Possibilities for a Better Enzyme. *Annual Review of Plant Biology*, 53(Volume 53, 2002):449–475. Publisher: Annual Reviews.
- Tello-Ruiz, M. K., Jaiswal, P., and Ware, D. (2022). Gramene: A Resource for Comparative Analysis of Plants Genomes and Pathways. In Edwards, D., editor, *Plant Bioinformatics*, volume 2443, pages 101–131. Springer US, New York, NY. Series Title: Methods in Molecular Biology.

- Teng, S., Madej, T., Panchenko, A., and Alexov, E. (2009). Modeling Effects of Human Single Nucleotide Polymorphisms on Protein-Protein Interactions. *Biophysical Journal*, 96(6):2178–2188. Publisher: Elsevier.
- The Angiosperm Phylogeny Group, Chase, M. W., Christenhusz, M. J. M., Fay, M. F., Byng, J. W., Judd, W. S., Soltis, D. E., Mabberley, D. J., Sennikov, A. N., Soltis, P. S., and Stevens, P. F. (2016). An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG IV. *Botanical Journal of the Linnean Society*, 181(1):1–20.
- The UniProt Consortium (2015). UniProt: a hub for protein information. *Nucleic Acids Research*, 43(D1):D204–D212.
- The UniProt Consortium (2023). UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Research*, 51(D1):D523–D531.
- Thomas, P. D., Campbell, M. J., Kejariwal, A., Mi, H., Karlak, B., Daverman, R., Diemer, K., Muruganujan, A., and Narechania, A. (2003). PANTHER: A library of protein families and subfamilies indexed by function. *Genome Research*, 13(9):2129–2141.
- Thornton, J. W. (2004). Resurrecting ancient genes: experimental analysis of extinct molecules. *Nature Reviews Genetics*, 5(5):366–375.
- Thumuluri, V., Almagro Armenteros, J. J., Johansen, A., Nielsen, H., and Winther, O. (2022). DeepLoc 2.0: multi-label subcellular localization prediction using protein language models. *Nucleic Acids Research*, 50(W1):W228–W234.
- Tian, W. and Skolnick, J. (2003). How Well is Enzyme Function Conserved as a Function of Pairwise Sequence Identity? *Journal of Molecular Biology*, 333(4):863–882.
- Trinh, M. D. L. and Masuda, S. (2022). Chloroplast pH Homeostasis for the Regulation of Photosynthesis. *Frontiers in Plant Science*, 13. Publisher: Frontiers.
- Tunyasuvunakool, K., Adler, J., Wu, Z., Green, T., Zielinski, M., Židek, A., Bridgland, A., Cowie, A., Meyer, C., Laydon, A., Velankar, S., Kleywegt, G. J., Bateman, A., Evans, R., Pritzel, A., Figurnov, M., Ronneberger, O., Bates, R., Kohl, S. A. A., Potapenko, A., Ballard, A. J., Romera-Paredes, B., Nikolov, S., Jain, R., Clancy, E., Reiman, D., Petersen, S., Senior, A. W., Kavukcuoglu, K., Birney, E., Kohli, P., Jumper, J., and Hassabis, D. (2021). Highly accurate protein structure prediction for the human proteome. *Nature*, 596(7873):590–596. Publisher: Nature Publishing Group.

- Van de Peer, Y., Mizrachi, E., and Marchal, K. (2017). The evolutionary significance of polyploidy. *Nature Reviews Genetics*, 18(7):411–424. Publisher: Nature Publishing Group.
- van Kempen, M., Kim, S. S., Tumescheit, C., Mirdita, M., Lee, J., Gilchrist, C. L. M., Söding, J., and Steinegger, M. (2024). Fast and accurate protein structure search with Foldseek. *Nature Biotechnology*, 42(2):243–246. Publisher: Nature Publishing Group.
- Varadi, M., Bertoni, D., Magana, P., Paramval, U., Pidruchna, I., Radhakrishnan, M., Tsenkov, M., Nair, S., Mirdita, M., Yeo, J., Kovalevskiy, O., Tunyasuvunakool, K., Laydon, A., Žídek, A., Tomlinson, H., Hariharan, D., Abrahamson, J., Green, T., Jumper, J., Birney, E., Steinegger, M., Hassabis, D., and Velankar, S. (2024). AlphaFold Protein Structure Database in 2024: providing structure coverage for over 214 million protein sequences. *Nucleic Acids Research*, 52(D1):D368–D375.
- Wade, R. C., Gabdoulline, R. R., and De Rienzo, F. (2001). Protein interaction property similarity analysis. *International Journal of Quantum Chemistry*, 83(3-4):122–127. _eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1002/qua.1204>.
- Wade, R. C., Gabdoulline, R. R., Lüdemann, S. K., and Lounnas, V. (1998). Electrostatic steering and ionic tethering in enzyme–ligand binding: Insights from simulations. *Proceedings of the National Academy of Sciences*, 95(11):5942–5949. Publisher: Proceedings of the National Academy of Sciences.
- Wang, L. G., Lam, T. T. Y., Xu, S., Dai, Z., Zhou, L., Feng, T., Guo, P., Dunn, C. W., Jones, B. R., Bradley, T., Zhu, H., Guan, Y., Jiang, Y., and Yu, G. (2020). Treeio: An R Package for Phylogenetic Tree Input and Output with Richly Annotated and Associated Data. *Molecular Biology and Evolution*, 37(2):599–603.
- Warshel, A. (1998). Electrostatic Origin of the Catalytic Power of Enzymes and the Role of Preorganized Active Sites *. *Journal of Biological Chemistry*, 273(42):27035–27038. Publisher: Elsevier.
- Warshel, A., Sharma, P. K., Kato, M., Xiang, Y., Liu, H., and Olsson, M. H. M. (2006). Electrostatic Basis for Enzyme Catalysis. *Chemical Reviews*, 106(8):3210–3235. Publisher: American Chemical Society.
- Waskom, M. (2021). Seaborn: Statistical Data Visualization. *Journal of Open Source Software*, 6(60):3021.

- Werck-Reichhart, D. and Feyereisen, R. (2000). Cytochromes P450: a success story. *Genome Biology*, 1(6):reviews3003.1.
- Werdan, K., Heldt, H. W., and Milovancev, M. (1975). The role of pH in the regulation of carbon fixation in the chloroplast stroma. Studies on CO₂ fixation in the light and dark. *Biochimica et Biophysica Acta (BBA) - Bioenergetics*, 396(2):276–292.
- Whisstock, J. C. and Lesk, A. M. (2003). Prediction of protein function from protein sequence and structure. *Quarterly reviews of biophysics*, 36(3):307–340. Publisher: Cambridge University Press.
- Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T., Miller, E., Bache, S., Müller, K., Ooms, J., Robinson, D., Seidel, D., Spinu, V., Takahashi, K., Vaughan, D., Wilke, C., Woo, K., and Yutani, H. (2019). Welcome to the Tidyverse. *Journal of Open Source Software*, 4(43):1686.
- Wierenga, R. (2001). The TIM-barrel fold: a versatile framework for efficient enzymes. *FEBS Letters*, 492(3):193–198.
_eprint: <https://febs.onlinelibrary.wiley.com/doi/pdf/10.1016/S0014-5793%2801%2902236-0>.
- Wilson, P. B., Estavillo, G. M., Field, K. J., Pornsiriwong, W., Carroll, A. J., Howell, K. A., Woo, N. S., Lake, J. A., Smith, S. M., Millar, A. H., Caemmerer, S. V., and Pogson, B. J. (2009). The nucleotidase/phosphatase SAL1 is a negative regulator of drought tolerance in Arabidopsis. *The Plant Journal*, 58(2):299–317.
- Winter, D., Vinegar, B., Nahal, H., Ammar, R., Wilson, G. V., and Provart, N. J. (2007). An “Electronic Fluorescent Pictograph” Browser for Exploring and Analyzing Large-Scale Biological Data Sets. *PLOS ONE*, 2(8):e718. Publisher: Public Library of Science.
- Wolfenden, R. and Snider, M. J. (2001). The Depth of Chemical Time and the Power of Enzymes as Catalysts. *Accounts of Chemical Research*, 34(12):938–945. Publisher: American Chemical Society.
- Woolfson, D. N. (2023). Understanding a protein fold: The physics, chemistry, and biology of α -helical coiled coils. *Journal of Biological Chemistry*, 299(4):104579.

- Wudarski, J., Egger, B., Ramm, S. A., Schärer, L., Ladurner, P., Zadesenets, K. S., Rubtsov, N. B., Mouton, S., and Berezikov, E. (2020). The free-living flatworm *Macrostomum lignano*. *EvoDevo*, 11(1):5.
- Xiang, S., Cooper-Morgan, A., Jiao, X., Kiledjian, M., Manley, J. L., and Tong, L. (2009). Structure and function of the 5'→3' exoribonuclease Rat1 and its activating partner Rai1. *Nature*, 458(7239):784–788. Publisher: Nature Publishing Group.
- Xiao, C., Li, K., Hua, J., He, Z., Zhang, F., Li, Q., Zhang, H., Yang, L., Pan, S., Cai, Z., Yu, Z., Wong, K.-B., and Xia, Y. (2023). Arabidopsis DXO1 activates RNMT1 to methylate the mRNA guanosine cap. *Nature Communications*, 14(1):202.
- Yang, Y., Ma, J., Yang, Y., Zhang, X., Wang, Y., Yang, L., and Sun, M. (2014). Yeast 3',5'-bisphosphate nucleotidase: An affinity tag for protein purification. *Protein Expression and Purification*, 97:81–87.
- Yoon, H. S., Grant, J., Tekle, Y. I., Wu, M., Chaon, B. C., Cole, J. C., Logsdon, J. M., Patterson, D. J., Bhattacharya, D., and Katz, L. A. (2008). Broadly sampled multigene trees of eukaryotes. *BMC Evolutionary Biology*, 8(1):14.
- Young, A. D. and Gillung, J. P. (2020). Phylogenomics — principles, opportunities and pitfalls of big-data phylogenetics. *Systematic Entomology*, 45(2):225–247. _eprint: <https://resjournals.onlinelibrary.wiley.com/doi/pdf/10.1111/syen.12406>.
- Yu, G., Smith, D. K., Zhu, H., Guan, Y., and Lam, T. T. Y. (2017). Ggtree: an R Package for Visualization and Annotation of Phylogenetic Trees With Their Covariates and Other Associated Data. *Methods in Ecology and Evolution*, 8(1):28–36.
- Yu, T., Cui, H., Li, J. C., Luo, Y., Jiang, G., and Zhao, H. (2023). Enzyme function prediction using contrastive learning. *Science*, 379(6639):1358–1363. Publisher: American Association for the Advancement of Science.
- Yu, X., Willmann, M. R., Vandivier, L. E., Trefely, S., Kramer, M. C., Shapiro, J., Guo, R., Lyons, E., Snyder, N. W., and Gregory, B. D. (2021). Messenger RNA 5' NAD⁺ Capping Is a Dynamic Regulatory Epitranscriptome Mark That Is Required for Proper Response to Abscissic Acid in Arabidopsis. *Developmental Cell*, 56(1):125–140.e6.

- Zakrzewska-Placzek, M., Golisz-Mocydlarz, A., Kwasnik, A., Krzyszton, M., Niedzwiecka, K., and Kufel, J. (2025). Defective Processing of Cytoplasmic and Chloroplast Ribosomal RNA in the Absence of Arabidopsis DXO1. *Plant, Cell & Environment*, 48(6):4227–4244.
- Zallot, R., Oberg, N., and Gerlt, J. A. (2019). The EFI Web Resource for Genomic Enzymology Tools: Leveraging Protein, Genome, and Metagenome Databases to Discover Novel Enzymes and Metabolic Pathways. *Biochemistry*, 58(41):4169–4182. Publisher: American Chemical Society.
- Zhang, J., Jin, H., Chen, Y., Jiang, Y., Gu, L., Lin, G., Lin, C., and Wang, Q. (2024). The eukaryotic translation initiation factor eIF4E regulates flowering and circadian rhythm in Arabidopsis. *The Plant Journal*, 120(1):123–138. [_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/tpj.16975](https://onlinelibrary.wiley.com/doi/pdf/10.1111/tpj.16975).
- Zhao, C., Wang, Y., Chan, K. X., Marchant, D. B., Franks, P. J., Randall, D., Tee, E. E., Chen, G., Ramesh, S., Phua, S. Y., Zhang, B., Hills, A., Dai, F., Xue, D., Gilliam, M., Tyerman, S., Nevo, E., Wu, F., Zhang, G., Wong, G. K. S., Leebens-Mack, J. H., Melkonian, M., Blatt, M. R., Soltis, P. S., Soltis, D. E., Pogson, B. J., and Chen, Z. H. (2019). Evolution of chloroplast retrograde signaling facilitates green plant adaptation to land. *Proceedings of the National Academy of Sciences of the United States of America*, 116(11):5015–5020.
- Zhou, H.-X. and Pang, X. (2018). Electrostatic Interactions in Protein Structure, Folding, Binding, and Condensation. *Chemical Reviews*, 118(4):1691–1741. Publisher: American Chemical Society.
- Zhou, X., Rao, S., Wrightstone, E., Sun, T., Lui, A. C. W., Welsch, R., and Li, L. (2022). Phytoene Synthase: The Key Rate-Limiting Enzyme of Carotenoid Biosynthesis in Plants. *Frontiers in Plant Science*, 13. Publisher: Frontiers.
- Zhou, X., Welsch, R., Yang, Y., Álvarez, D., Riediger, M., Yuan, H., Fish, T., Liu, J., Thannhauser, T. W., and Li, L. (2015). Arabidopsis OR proteins are the major posttranscriptional regulators of phytoene synthase in controlling carotenoid biosynthesis. *Proceedings of the National Academy of Sciences*, 112(11):3558–3563. Publisher: Proceedings of the National Academy of Sciences.
- Zimmermann, P., Hirsch-Hoffmann, M., Hennig, L., and Gruissem, W. (2004). GENEVESTIGATOR. Arabidopsis Microarray Database and Analysis Toolbox. *Plant Physiology*, 136(1):2621–2632.